Coupling transport and chemistry: numerics, analysis and applications

Mitchell, Sarah Louise

Award date:
2003

Awarding institution:
University of Bath

Coupling Transport and Chemistry: Numerics, Analysis and Applications

submitted by

Sarah Louise Mitchell

for the degree of PhD

of the

University of Bath

2003

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

To my parents

Coupling Transport and Chemistry: Numerics, Analysis and Applications

or

A Tale of Tiny Wiggles

Sarah Louise Mitchell

Abstract

This work is concerned with the analysis and numerical approximation of systems of differential equations which describe the transport of chemicals in groundwater flow.

These equations have the general form of conservation laws (representing the groundwater transport) with source terms (representing the chemical reactions). In typical problems the reaction rates vary widely, and at least some will be on a much smaller time scale than that implied by the advection velocities. The result is the phenomenon of reduced (or retarded) speed whereby the transport of pollutants is much slower than the flow of the groundwater.

Contributions in the Thesis have been made in both analysis and numerics. Classical PDE techniques are applied to a simple reactive transport model which enables us to derive, from first principles, theoretical results establishing exponential decay of the solution profile as it evolves in time. This result is used to obtain bounds on the reduced speed by integrating the conservation law over a given domain.

In the design of numerical methods we wish to take time-steps guided by this reduced speed. However, we show that the standard practice of using operator splitting methods with explicit time-stepping, to maintain stability, requires much smaller time-steps than might be expected from a knowledge of the retardation behaviour.

We apply a combination of the box scheme and trapezoidal scheme (box-trap scheme), both in one and two space dimensions, to both simple problems and more complex systems. The box scheme is the standard numerical method for river-flow modelling which also exhibits numerically the phenomenon of reduced speed. It is very compact and so easily accommodates rapid changes in the parameters that can arise. Moreover, the scheme is unconditionally stable which allows us to choose the most appropriate time-step to best represent the speed and accuracy of the solution. The box-trap scheme is shown to robustly handle the different speeds that can occur in larger systems.

The main disadvantage of the scheme is the presence of spurious oscillations in the numerical solution. We show that these can be damped by introducing a time-weighting of the spatial derivatives and this does not noticeably affect the accuracy. A technique called the modified equation analysis is used to give considerable insight into both the numerical scheme and original model. The resulting expansions enable us to control and damp the oscillations. A new approach leads to the modified equation expansion of the oscillatory part of the solution which means we can predict the exact position of the oscillations.

We have also successively adapted the box scheme to numerically solve nonlinear conservation laws with shocks which has applications to coupled reactive transport models.

Acknowledgements

I would like to begin by thanking my supervisors Alastair Spence and Bill Morton for their contribution to this Thesis. Their guidance, advice, enthusiasm, humour and encouragement over the past three years has been invaluable. I will miss working with them both very much. I thank Alastair for giving me the opportunity to work on such an interesting project and for all our discussions on mathematics and many other topics. I also thank Bill who, although not officially my supervisor, has become an essential part of this project. I am very grateful for his endless patience and time, his initiative and unforgettable advice ("you need to take a bath Sarah" will always stick in my memory!) I should thank First Great Western trains for all their delays: many an email from Bill has begun "I've had an idea on the train"!

A big thank you must go to my family, especially my parents. They have provided great support, encouragement and advice throughout all my education and I would not have achieved this without them. I also wish to mention my Granny, who sadly passed away earlier this year, but always showed a great interest in my work and would have loved to have seen my final bound Thesis.

Furthermore, I would like to thank Matt Piggott for all his academic help in my time at Bath. His patience and support are very much appreciated and his vast background knowledge has been both helpful and inspiring. Likewise, thanks to JF Williams for his time, advice and interest in my work and to Jörg Berns-Müller for endless help with Matlab and LaTeX.

In addition, I thank the EPSRC for their financial support and the Department of Mathematical Sciences, especially the Numerical Analysis group and the administration staff (Mark Willis in particular), for providing a great working environment.

My last thanks go to all my friends in Bath who have made my time as a postgrad so enjoyable. I especially mention Marc, Matt and Bob: I will always remember the conversations and endless coffee breaks in the little kitchen on level 2. Thanks to Marianne for being an excellent friend and a great laugh and keeping me sane (just!) in these last few months of writing up: KBO girl, it will all be worth it I promise! Also, thanks to Darran, Tracey and Richard, who I've really enjoyed living with throughout the last three years; Jon, for listening to my moans and ramblings whilst running; Ivana, for her friendship which began after our joint humiliation by Victor in Mathematical Methods I; Team 1W 3.8, for being such nice office mates; and lastly, the "gaggle", for some great nights out in Bath.

Finally, a quote which I decided not to put on its own page but is very apt: "This report, by its very length, defends itself against the risk of being read" (Winston Churchill).

Contents

Chapter 1

Introduction

1.1 Problems and models

This Thesis is concerned with the analysis of models that describe the coupling of transport and chemistry in groundwater flow; that is, models representing how concentrations of chemical pollutants react whilst being transported through the groundwater. They can be carried far from their source and affect the surrounding environment. Examples of chemical interactions with transport arise in problems such as the migration of chemical contaminants that may threaten public groundwater supplies or the long term isolation of radioactive waste. We assume that the chemical pollutants react in the water, both with each other and the rock; and so are adsorbed into the rock and then desorbed back into the water at a later time. This means that a complete reactive groundwater transport model must account for both of these processes. The general philosophy of multicomponent reactive modelling and an overview of the mathematical problem formulation and the nature of the chemical reactions is discussed in (Rubin 1983).

Systems of this type often exhibit *retardation*, i.e. the chemical reactions produce a transport speed which is slower than the advection speed itself. This occurs when the chemical reactions are fast compared with the advection rate. Then the chemical pollutant will not move at the advected speed, but at a reduced speed. This is well known in the literature for other applications as well as reactive transport. For example, (Whitham 1974, page 28) formulated the equations for exchange processes between a solid bed and a fluid flowing over it. Lower order waves are shown to move at a slower speed than the fluid flow. Similarly, this is confirmed in (Rhee, Aris & Amundson 1986, page 160) by exactly solving a simple model describing the same process. The phenomena of retardation is also mentioned in various papers from water resources journals including (Zysset, Stauffer & Dracos 1994, pages 2222, 2224), (Herzer & Kinzelback 1989, page 121) and (Yeh & Gwo 1990, page 419). When numerically

solving such systems by explicit time-stepping, it is preferable to take time-steps guided by this retarded speed rather than the advection speed. Many numerical methods will often require, to maintain stability, much smaller time-steps than might be expected from a knowledge of the retardation behaviour.

In situations where there are numerous chemical pollutants it is very likely that there will be varying retardation speeds in the system, as discussed in (Yeh & Gwo 1990, page 425). Some of the speeds will be close to the advection speed whilst others will be severely reduced. This can cause difficulties with the numerical method as it will not be clear how to choose the time-step for these systems; it may well lead to stability restrictions (as shown in the internal report (Budd, Carey, Graham & Spence 1997, page 19)). We will propose a numerical method which is robust enough to handle these varying speeds and will be computationally accurate for both the largest and smallest speed that arises.

Another interesting feature of these models is the presence of diffusion. Since the models do not have explicit diffusive terms, this property is not generally obvious from the model equations. A short injection of chemical pollutants into the water at a particular point in space will be diffused and become more spread out as it moves through the water: it is this decay that we wish to observe at later points in space.

The phenomena described above make these systems interesting and challenging to study, both analytically and numerically. We will perform various analyses on simple models which describe coupling transport and chemistry to show these features.

These systems can be compared with flood waves in long rivers. Lighthill & Whitham (1955) wrote an influential paper about the theory of a distinctive type of wave motion, which arises in any one-dimensional flow problem. This class of wave motions, known as *kinematic waves*, are physically quite distinct from the classical wave motions encountered in dynamical systems, known as *dynamic waves*. Kinematic waves exist if, to a sufficient approximation, there is a functional relationship between the flow and the concentration. The wave property then follows directly from the resulting equation of continuity. In contrast, dynamic waves depend on Newton's second law of motion, together with some assumption on the stress (e.g. for gravity waves the assumption relates a stress to a displacement). An important difference is that kinematic waves possess only one wave velocity at each point, and therefore the flow remains constant on each kinematic wave, whereas dynamic waves possess at least two wave velocities. Kinematic waves are not dispersive, but they suffer change of form due to nonlinearity. In Section 3 of the paper it is shown that, in the case of flood waves, it is possible for both kinematic and dynamic waves to occur together. However, dynamic waves have a much higher wave velocity and also a rapid decay. Hence, although some disturbance is

sent downstream at the ordinary wave velocity for dynamic waves, at any considerable distance downstream this is negligible and so the main wave will be kinematic and moves at a much slower velocity. They then assume a dominant role.

This shows that flood waves have some of the same features as coupled transport and reaction systems. If we think of the movement of chemicals in the discussion above as being one dimensional then the coupled transport and reaction equations have three key terms: the time variation, the space variation (which together give the advection) and the local term (which is the reaction in this case). For flood waves, the local term is the bed friction (i.e. the frictional force of the river bed). The balance between the bed friction and the spatial derivative dominates the system and also gives a different speed, also discussed in (Whitham 1974, page 82).

1.2 Numerical methods

A major task facing geochemists today is the demand for reliable modelling of groundwater transport of chemical pollutants. One has to decide whether to formulate the chemical reactions kinetically or to assume a local equilibrium state. The most general description of the chemical process is a kinetic formulation in which the linear partial differential equations of transport are coupled with the nonlinear system of ordinary differential equations describing the kinetic development. In certain cases the chemical process can be modelled by assuming them to be in local chemical equilibrium; this leads to the equilibrium formulation where the linear partial differential equations of transport are coupled with the nonlinear system of algebraic equations describing a chemical equilibrium state. This assumption is only valid when the chemical reactions are so fast compared with the rate of transport that they appear to react at a fixed point in time.

For both these formulations there are two approaches to the numerical solution which are described in water resources journals, (Zysset et al. 1994), (Herzer & Kinzelback 1989) amongst others. These are known as one-step methods (where the chemical and transport processes are solved simultaneously), and two-step methods (where the processes are separated). This allows the reactions to be decoupled from the transport equations so that an operator splitting technique can be applied. In the literature review in Section 4 we will explain these procedures in more detail.

Although flood waves have some of the same characteristics as reactive transport models, different approaches are commonly used to solve these two processes numerically. Firstly, reactive transport models concentrate on the chemical reaction and tend to assume that the reactions satisfy local equilibrium conditions (and so the chemistry is conserved). This is because a large amount of data exists about equilibrium parameters

in the chemical literature; also, the resulting algebraic equation system is faster and easier to solve. In contrast, the reactive transport through the rock is very complicated and so it is not easy to predict how the chemicals are adsorbed and then desorbed back into the water. There is a well known methodology in equilibrium chemistry and good methods exist for approximating this process numerically. As noted in (Friedly & Rubin 1992, page 1935), coupling geochemistry equilibrium models with transport models is still relatively new. Hence researchers in the field wish to use the existing algorithms and combine them with a transport step to cover the whole process. Secondly, for flood waves, more information is known about the flow of water in a channel than the bed friction (which is analogous to the reaction term). Hence numerical methods to approximate these problems concentrate on conserving the water.

In this Thesis we follow the technique used in the study of river modelling and assume the amount of concentration of chemical pollutants in the water and rock are conserved; we do not want to be restricted to assuming chemical equilibrium. We consider an implicit finite difference scheme called the *Preissman four-point scheme*, otherwise known as the *box scheme* (Cunge & Holly Jr 1980, page 65). It is based on integral relationships and we will show this explicitly by deriving the scheme in Chapter 3. This is the chosen method for hydraulics engineers in the study of river modelling (Abbott 1979, page 188) where the flood wave flows at a slower speed than the fluid velocity. An example of this is the St. Venant equations (Cunge & Holly Jr 1980, page 8) which are generally regarded as providing a valid model for the study of one dimensional channel flows. The scheme is second order accurate in space and time and its compactness easily accommodates rapid changes in the parameters. Its unconditional stability allows the use of comparatively large time-steps which will enable us to choose the most appropriate to best represent the speed of the solution. The disadvantage of the box scheme is the spurious modes which can become very prominent for coupled systems.

Preissmann developed the box scheme to deal with unsteady flow in open channels in 1961 (as discussed in (Cunge & Holly Jr 1980, page 66)). We will apply this method to reactive transport models, both for simple model problems and larger systems with several chemical pollutants. We will show that the box scheme can handle the varying retardation speeds that can arise whilst remaining computationally accurate. A simple explicit finite difference scheme will be considered as a comparison and we will see that the stability restriction becomes severe (and so requiring very small time-steps) when the retardation speed is much less than the advection speed.

However, the box scheme contains a spurious solution which causes non-physical oscillations in the numerical approximation (and is typical of second and higher order methods). For the linear advection equation this persists for all time with no damping (because there is no damping of the modes). These oscillations are worse than, say

those in the Lax-Wendroff scheme, but can be avoided by using a time-weighting of the spatial differences. In fact, the Preissmann four-point scheme described in (Cunge & Holly Jr 1980, page 65) already has this parameter included; its use is standard practice by hydraulics engineers. The effects of using a weighting parameter will be examined in Chapter 3. Our task is to use the weighting to reduce these oscillatory modes and formulate a discrete system of equations which can be applied to many models.

We also have to be careful that we try to avoid numerical diffusion wherever possible. (Cunge & Holly Jr 1980, page 325) suggest the following measures to choose Δx and Δt which will minimise the effect of the artificial diffusion:

1. use Δx as small as possible within the time and cost limits of the study;
2. use Δt as large as possible without losing resolution in the description of the dispersion process;
3. choose Δt and Δx such that $\Delta x \approx V\Delta t/n$, where V is the advection speed and n is a positive integer. So, if there is little retardation in the model, n would be one and this would increase as the difference between the retardation speed and advection speed becomes larger.

Jennings, Kirkner & Theis (1982, page 1089) state that many engineers working in this field would like the interaction chemistry to be posed independently of the transport equations. Even though the box scheme does not follow this approach, the discretised equations can be formulated in a similar way and introduce a minimum modification of the equilibrium solver. Then the numerical code in existence for this solver could easily be adapted to solve the box scheme applied to the full system (i.e with the kinetic formulation for the chemical process). Hence, instead of solving the equilibrium formulation using an existing two-step method, we would solve the kinetic formulation using an unconditionally stable one-step method (i.e. the box scheme); and we know the latter will give much more accurate results (Mitchell, Morton & Spence 2003*a*).

1.3 Derivation of simple models

Suppose a single chemical pollutant is transported by the groundwater through a rock. In a simple model of transport considered here, we neglect the dispersive effects of heat conduction, diffusion or viscosity and only consider convective transport. At a certain time level the pollutant is adsorbed into the rock and at a later time is desorbed from the rock back into the water. Let a denote the concentration of the chemical pollutant in the water, b the concentration of the chemical pollutant in the rock and A the cross-sectional area of the rock. If x is the direction of flow and V is its velocity then we

have a simple mass balance over a volume $A\triangle x$

$$(VaA)_{x,t} - (VaA)_{x+\triangle x,t} = A\triangle x \left\{ \frac{\partial}{\partial t}[a+b] \right\}_{\bar{x},t}, \quad x < \bar{x} < x + \triangle x$$

where the left hand side is the rate of flow of species into the infinitesimal volume $A\triangle x$ minus the rate of outflow, and the right hand side is the rate of accumulation of solute into the volume $A\triangle x$.

Assuming V and A are constant and taking the limit as $\triangle x \to 0$, we obtain

$$V\frac{\partial a}{\partial x} + \frac{\partial a}{\partial t} + \frac{\partial b}{\partial t} = 0, \tag{1.1}$$

or, in alternative notation

$$a_t + b_t + Va_x = 0. \tag{1.2}$$

This is our basic conservation equation. It involves no assumption about the mechanism of adsorption and so, assuming there is a finite transfer of solute from the water to the adsorbent, we must couple this with a reaction equation (often determined empirically) which describes the relationship between b and a. In general terms we write

$$b_t = -f(a,b), \tag{1.3}$$

with the following assumptions made on f:

$$\frac{\partial f}{\partial a} \leq -\lambda, \quad \frac{\partial f}{\partial b} \geq \mu, \tag{1.4}$$

where λ and μ are strictly positive constants. Hence our basic model describing a single chemical pollutant travelling through the groundwater (and being adsorbed and desorbed from a rock to the water) is given by (1.2) and (1.4). This can be written as

$$a_t + Va_x = f(a,b) \tag{1.5}$$

$$b_t = -f(a,b). \tag{1.6}$$

From now on (1.5) will be known as the *Transport equation* and (1.6) as the *Reaction equation.*

Guenther & Lee (1988, pages 210-212) discuss different ways of describing the reaction term. The simplest situation occurs when the concentration b of chemicals adsorbed in the rock is proportional to the solute concentration a. Then the equation (1.6) is no longer needed, but instead we have $b = \alpha a$, for some constant $\alpha > 0$. Inserting this

into (1.2) gives the conservation equation

$$a_t + \frac{V}{1+\alpha} a_x = 0, \tag{1.7}$$

which immediately shows the retardation in the system since $V/(1+\alpha) < V$.

We now turn to a more realistic assumption: suppose there is a constant maximum concentration $\tilde{b}$ at which the rock becomes saturated and that the adsorption rate is proportional to the difference $\tilde{b} - b$. Then (1.6) can be written specifically as

$$b_t = \alpha(\tilde{b} - b), \tag{1.8}$$

again for some constant $\alpha > 0$. If initially there is no chemical present in the rock, $b(x, 0) = 0$, and we can integrate (1.8) to give

$$b(x,t) = \tilde{b}(1 - e^{-\alpha t}). \tag{1.9}$$

Substituting this expression for b into (1.2) yields

$$a_t + V a_x + \alpha \tilde{b} e^{-\alpha t} = 0. \tag{1.10}$$

Of course, the parameter α in (1.8) will in practice depend on the concentration a. The simplest assumption taking this dependence into account leads to the reaction equation

$$b_t = \beta a(\tilde{b} - b), \tag{1.11}$$

where $\beta > 0$ is constant. The reaction term in (1.11) is nonlinear and so less amenable to analysis.

We now describe three models which will be the basis of our study.

1.3.1 The Linear Model

The right hand side of (1.11) can be linearised to obtain a model which we can use to perform some analysis (i.e. $\beta a(\tilde{b} - b) \approx \beta \tilde{b} a - \beta a_0 b$). The adsorption rate should increase as the concentration a increases; that is, b_t should be an increasing function of a. Similarly, as more and more chemical is adsorbed, the ability of the solid to adsorb the chemical will decrease (these are precisely the conditions given in (1.4)). So, b_t will be a decreasing function of b. The simplest model with these characteristics is a linear reaction term

$$b_t = \lambda a - \mu b, \tag{1.12}$$

where λ is the proportion of chemical pollutants adsorbed from the water to the rock and μ is the proportion of chemical pollutants desorbed into the water from the rock (with $\lambda,\ \mu > 0$). Hence the right hand side is the finite rate of transfer of the solute from the groundwater to the adsorbent (i.e. the rock). Physically, it is obvious that we require $\lambda > \mu$, and often we have that $\lambda \gg \mu \gg 1$: the rock will more easily be able to adsorb a chemical pollutant than to desorb it back into the water. If this is the case and $V = O(1)$ then we say that the reaction term is *stiff*. Let us rewrite (1.2) and (1.12) in the following form:

$$a_t + Va_x = -\lambda a + \mu b \tag{1.13}$$

$$b_t = \lambda a - \mu b. \tag{1.14}$$

We call (1.13) and (1.14) the *Linear Model*. Using the terminology from (1.3), $f(a,b) = -\lambda a + \mu b$ and so the conditions from (1.4) are satisfied. Note that this model is physically realistic only for very dilute solutions. Also, the effects of changes in temperature are ignored.

1.3.2 The Langmuir Model

Another form the rate of adsorption might assume is

$$b_t = \lambda a(B - b) - \mu b. \tag{1.15}$$

Here λ and μ are still the same rate constants described above and B is the saturated concentration of b in the rock. Thus the rate of which a is adsorbed into the rock slows as b increases. This is more realistic in chromatography when saturation occurs. Hence $f(a,b) = -\lambda a(B - b) + \mu b$ and so the conditions from (1.4) are satisfied provided B and b are scaled appropriately so that $B - b \leq 1$ (and provided we assume $a \geq 0$).

Under equilibrium conditions this rate law corresponds to what is known as the Langmuir Isotherm (Rhee et al. 1986, page 35), i.e.

$$b = B\frac{Ka}{1 + Ka}, \qquad K = \frac{\lambda}{\mu}. \tag{1.16}$$

Hence we refer to (1.15), coupled with (1.2), as the *Langmuir Model*.

1.3.3 The Flushing-through Model

A recent area of research considered by AEA Technology Harwell (now SERCO Assurance) has been to study the evolution of the chemical environment in and around a nuclear waste repository. During groundwater transport, reactive solutes are subject to a variety of hydro-physical and chemical processes. These hydro-physical processes

include advection and diffusion and the chemical processes are time dependent. Combining the systems leads to a system of six partial differential equations with quadratic nonlinearities which is described in Chapter 5.

A particular case of interest is when one or more species are flushed through the system; and so there is not much retardation present in the system. This means the six equation model can be reduced to a three equation model which has two chemical pollutants in the water, a and c. The concentration a will be adsorbed into the rock (and so creating b) whilst the concentration c will simply be carried along through the water, only reacting with a. We think of this as "c being flushed through the system" and call the resulting equations the *Flushing-through Model.* This is given by

$$a_t + Va_x = -\lambda a + \mu b - \gamma a^2 + \delta c \tag{1.17}$$

$$b_t = \lambda a - \mu b \tag{1.18}$$

$$c_t + Vc_x = \gamma a^2 - \delta c, \tag{1.19}$$

where γ and δ are reaction constants. Concentrations a and c could travel at different speeds through the groundwater for various values of the parameters λ, μ, γ and δ. We will investigate how the box scheme copes with this phenomenon in Chapter 5.

1.3.4 The general problem

The above three model problems can be formulated in a more general way. Suppose $\mathbf{u_1}$ and $\mathbf{u_2}$ are state vectors of dimension p_1 and p_2 respectively and let d be the number of spatial dimensions in the system. Then the general form which describes chemically reacting flow is given by

$$(\mathbf{u_1})_t + \nabla \cdot \mathcal{F}(\mathbf{u_1}) = \mathbf{S_1}(\mathbf{u_1}, \mathbf{u_2}) \tag{1.20}$$

$$(\mathbf{u_2})_t = \mathbf{S_2}(\mathbf{u_1}, \mathbf{u_2}), \tag{1.21}$$

where $\mathbf{u_1}(\mathbf{x}, t) : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^{p_1}$, $\mathbf{u_2}(\mathbf{x}, t) : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^{p_2}$, the fluxes $\mathcal{F}$ represent the groundwater transport and the terms $\mathbf{S_1}$ and $\mathbf{S_2}$ represent the chemical reactions. In this Thesis we will mainly discuss the three model problems in which, for both the Linear and Langmuir Models, $\mathbf{u_1} = a$ and $\mathbf{u_2} = b$; and for the Flushing-through Model $\mathbf{u_1} = [a, c]^T$ and again $\mathbf{u_2} = b$. However, our methods (both analytical and numerical) will be applicable for the general model (1.20) and (1.21).

1.4 Literature review

The general problem (1.20) and (1.21) is an example of a hyperbolic system of conservation laws with relaxation. The majority of work in the numerical analysis literature

on this area uses operator splitting methods. We will discuss the relevant literature from the water resources journals and numerical analysis journals separately.

From reading the water resources journals it is clear that there are two main issues: the first is whether to formulate the chemical reactions kinetically or to assume a local equilibrium state. The majority of the literature assumes the latter but this is not always a good idea since, as (Friedly & Rubin 1992, page 1935) state, "there is a growing amount of experimental evidence suggesting that equilibrium is not always attained in subsurface transport. Kinetic rates of both chemical reactions and physical diffusion processes can be important". However, despite this, there is still little discussion of systems with kinetically controlled reactions, although (Molz, Widdowson & Benefield 1986), (Zysset et al. 1994) and (Wheeler & Dawson 1988) have done work on both formulations. The second issue is whether to use one or two-step solution methods. Herzer & Kinzelback (1989) give an overview of both solution methods if the system is assumed to be in chemical equilibrium. The task then consists of solving a combined system of PDEs and algebraic equations (AEs). In the one-step approach the sets of PDEs and AEs are solved simultaneously. This is mathematically more rigorous but the disadvantage is that there maybe high computational cost in evaluating the operational equations. It is far more common to use a two-step method.

In the two-step approach each time-step is split into a transport step involving the PDEs for the mass transport, and a reaction step which solves the set of AEs for the equilibrium chemistry. This can be done in several ways; Herzer & Kinzelback (1989) mention two possible schemes:

1. the chemical reaction rates are estimated from the initial concentrations of each time level (the reaction step). These are then inserted into the source terms of the transport equation;

2. the transport equations are solved setting the source terms to zero first (a pure transport step). This results in intermediate concentrations and the reaction rates and final concentrations are then computed (reaction step).

In all two-step methods an error is introduced from the decoupling of the transport and chemical equations. This can be reduced by iteration between the two steps. If the above examples are used without iteration then the second is preferable as it requires no additional stability conditions. If both these examples are iterated then they actually coincide and represent a predictor-corrector scheme for reactive multicomponent transport. A disadvantage of this approach is that the iterative procedure is not unconditionally stable due to the explicit nature of the two-step approach.

However, there are some advantages of two-step methods which show why it is a popular way to approximate reactive transport systems. It is highly flexible in switching

from one chemical system to another by simply replacing appropriate chemical reaction subroutines. Also, the concept can be applied to the kinetic approach. Herzer & Kinzelback (1989) concentrate on the second example described above and conclude that it is a powerful method but that the stability and rate of convergence is not unconditional. They depend on the chosen numerical approximation of the transport process and the size of the explicitly introduced source term. Hence, if the source terms are large the stability restriction will be severe and require very small time-steps.

Barry, Miller, Culligan & Bajracharya (1997) and Jennings et al. (1982) also outline the advantages of using splitting methods. Barry et al. (1997) states that coupled solute transport and reaction models are computationally demanding when multispecies, multidimensional simulations are considered. They suggest that split operator methods provide approximate solutions to the reactive solute transport problem that are both relatively efficient to compute and to construct. Jennings et al. (1982) claim that separating the two sets of equations has been used with some success and offers some attractive advantages. There are comprehensive computer codes in operation which are especially effective for solution equilibrium computations when the number of components is large. However, although they are very powerful, there are some disadvantages in their use. The routines can easily be non-convergent and, since the existing packages are very general, they are also very large which can slow down the computations.

In (Barry, Bajracharya & Miller 1996) a two-step split operator method is considered. The resulting error is proportional to Δt and so very small time-steps are required to ensure accurate solutions. Consequently, the authors also describe a method which enables accurate solutions to be calculated more efficiently but maintains the separation of the transport and reaction steps. This method uses LU factorisation and so is limited to transport problems where this can be applied. Barry et al. (1997) also extend their work to multispecies groundwater chemical transport models. Here the standard two-step method is considered, along with an alternating split operator scheme. The order of operations is switched at succeeding time-steps which removes the splitting error for linear reactions, but not in nonlinear cases.

In (Zysset et al. 1994), where both the kinetic and equilibrium formulations are considered, the work uses a sequential two-step method for the former case and an iterative two-step method for the latter. Comparisons are also made between sequential and iterative two-step methods in (Walter, Frind, Blowes, Ptacek & Molson 1994*a*), although attention is restricted to the equilibrium formulation. The authors suggest that the most efficient approach is to use a sequential two-step method, without iteration between the physical and chemical processes, which limits the extent of the equation nonlinearity in the chemistry step.

Hence, in the study of coupled reactive transport models the main approaches to solving these systems numerically has been to use two-step splitting methods and to immediately assume that the reactions are in equilibrium. However, we follow a different approach because we do not wish to be restricted to assuming an equilibrium state or using very small time-steps to ensure the stability conditions are satisfied.

We now discuss the literature in numerical analysis journals for the numerical solution of hyperbolic systems of conservation laws with relaxation. They can be used to describe many physical problems, for example flood waves in long rivers. There are other well known examples including viscoelasticity (the memory effects are modelled as relaxation), magnetohydroynamics and traffic flow.

A system of hyperbolic conservation laws with relaxation consists of one or more conservation laws coupled with one or more rate equations (Pember 1993*a*, page 1294). Each rate equation governs the time evolution of a nonconserved quantity. A rate equation has an associated timescale called a relaxation time, which determines how quickly the nonconserved quantity approaches its equilibrium state. The limit of the system as the relaxation times vanish is a smaller system in which the rate equations are replaced by functions expressing the equilibrium value of the nonconserved quantity as a function of the conserved quantities. We assume the smaller system is also hyperbolic. In such systems the relaxation time may vary from order one to much less than unity. A system of conservation laws with relaxation is called *stiff* when at least one of its relaxation times is small compared to the timescale determined by the characteristic speeds of the system. These varying time scales can lead to numerical difficulties similar to stiff systems of ODEs. As (Pember 1993*a*) asks "can we obtain an accurate numerical solution of a stiff, hyperbolic system of conservation laws with relaxation using time and space increments governed solely by the non-stiff part of the system, i.e., without fully resolving the effect of the stiff source terms?"

As described in (Caflisch, Jin & Russo 1997, page 247), earlier work in this field concentrates on developing robust numerical schemes that handle the stiffness of the problem effectively and avoid spurious numerical solutions (which occur when the grid spacing under-resolves the small relaxation time). By "spurious solution" we mean a nonphysical numerical solution that bears no resemblance to the actual solution but satisfies the discrete equations (Pember 1993*a*, page 1294). However, these earlier schemes often did not have high-order accuracy uniformly with respect to the wide range of relaxation times.

Caflisch et al. (1997) have developed a class of numerical methods using implicit finite difference equations with uniform accuracy. A splitting scheme is used for the advection and reaction steps. High order flux in the convection step gives high spatial accuracy.

A second-order time discretisation can then be obtained either by using Richardson extrapolation or by a suitable combination of relaxation and convection steps. Cecchi, Redivo-Zaglia & Russo (1996) seek to develop a robust numerical scheme that will work uniformly for a wide range of relaxation rates. The authors extend this idea in (Caflisch et al. 1997) and improve the order of convergence of the time discretisation of these splitting schemes.

Colella, Majda & Roytburd (1986) and Pember (1993*b*) give accurate algorithms for gas dynamics which used a splitting technique, i.e. integration of the gasdynamic terms first and then integration of the appropriate ODE for the source term in an intermediate step. This decoupling can be done in an optimal way using Strang-type splitting (Strang 1968). Unfortunately, even this technique still introduces numerical errors and, in reality, the transport and chemistry are strongly coupled and cannot be separated. However, Strang splitting is second order accurate and so, at least for smooth solutions, LeVeque & Yee (1990) suggest that the interaction of different effects can be modelled adequately by a split method. There are advantages to splitting since high quality numerical methods have been developed both for systems of conservation laws and for stiff systems of ODEs and so these methods can be used directly.

Botchorishvili, Perthame & Vasseur (2003) consider *equilibrium schemes* for scalar conservation laws with stiff source terms. They compare this with finite volume schemes, which are well known to give unsatisfactory results because of the lack of accuracy on the equilibrium states. Source terms should be taken into account in the upwinding and discretised at the nodes of the grid. Numerically, equilibrium schemes involve solving the conservation law with no source term using an upwind finite volume scheme. This incorporates the discrete equilibrium states which are defined according to steady state equations (involving the source terms). Numerical tests in (Botchorishvili et al. 2003) show that this splitting scheme is far more accurate than the usual upwind method.

However, LeVeque & Yee (1990) show why splitting methods can be inadequate when applied to hyperbolic conservation laws with stiff source terms for discontinuous data. A scalar problem with three equilibrium states is considered and stable results are obtained which are free of oscillations but move at the wrong speed. It is claimed that this is purely due to the numerical method. The problem lies with the smearing of the discontinuity caused by the advection; this introduces intermediate states which are not in equilibrium. As soon as the non-equilibrium value appears, the source term turns on and immediately restores equilibrium. In the non-stiff case the results are reasonable; however, in the stiff case reducing the time-step is not enough to overcome this problem. The spatial resolution is as important as the temporal resolution and so it is necessary to consider alternatives to uniform finite difference methods. Although we will always assume that our source term has only one equilibrium state, LeVeque &

Yee (1990) show that their splitting methods are only really adequate in non-stiff cases (without refinement of the overall grid which can be expensive).

Unsplit numerical methods have been developed which should also be mentioned. Extensive work has been done in (Pember 1993*b*), (Papalexandris, Leonard & Dimotakis 1997) and (Jin & Levermore 1996) which use a variety of techniques including Roe's approximate Riemann solver and higher-order Godunov methods. These methods are very effective but are restrictive as they assume all the relaxation times are small. We wish to develop a method which works for a large range of these times (and thus for a system with varying speeds). Split numerical methods are better at dealing with this wide range of relaxation times but accuracy is lost and can give the wrong speed. Hence we take a different approach and consider the box scheme.

1.5 Outline of Thesis

We now give the outline of this Thesis and state which results are new in the field.

Chapter 2 focuses on the Linear Model and related models and some interesting analytical results are proved. We consider the limiting behaviour of such systems with stiff relaxation and use a time-asymptotic expansion to study the process of relaxation to its equilibrium states. A simplified system is derived which shows the reduced speed and we can see explicitly the presence of diffusion which is not immediately obvious from the original formulation. This is a well known technique in the study of hyperbolic conservation laws with relaxation (Liu 1987). We also apply the methods from (Lighthill & Whitham 1955), which considers flood movement in long rivers, to prove that the wave decays exponentially. An approximate form of the solution for large time can then be found and the retarded speed again deduced. Lastly in this Chapter we look at conservation properties of the general two equation model with nonlinear $f(a, b)$. It can be written as a second order equation in one variable and then well known techniques from the theory of linear second order PDEs (Garabedian 1964, pages 127-135) are used to obtain a solution in terms of definite integrals on a given domain. This leads to new results showing how the solution behaves at infinity and we can prove that the concentration of chemical pollutants will move at a slower speed than the speed of advection for a general nonlinear reaction term.

Chapter 3 is concerned with an analysis of the box scheme applied to linear problems. We begin by studying how the box scheme approximates a simple conservation law and discuss the spurious oscillatory solution that arises in the numerical solution for non-smooth data. A combination of the box scheme and the trapezoidal rule is then applied to the Linear Model. By using a time-weighting for the spatial differences we show that these oscillations can be avoided. A modified equation analysis (Warming

& Hyett 1974) gives us a greater understanding of these oscillations and leads to an expansion which matches the simplified system found using the asymptotic analysis. This is a new approach and we are able to use this expansion to give previously unseen results; in particular it can guide us to the best choice of the CFL number to reduce the oscillations and it can predict their exact position.

In Chapter 4 we apply the box scheme to nonlinear conservation laws with shocks. The box scheme is inadequate for computing discontinuous solutions of nonlinear conservation laws and is worse than other second order schemes, for example the Lax-Wendroff scheme. We derive the box scheme as a Petrov-Galerkin method using a piecewise constant test space and a piecewise linear trial space. With this viewpoint in mind we develop an algorithm which uses a different trial space across the cell containing the shock. This essentially post-processes the numerical solution to eliminate the oscillations before moving to the next time-step. When this correction is applied to the (inviscid) Burgers equation the oscillations are almost entirely eliminated. The box scheme has not been used in the past to solve nonlinear conservation laws in the presence of shocks but with this correction it is very effective. We will apply this to the Langmuir Model which, although does not have a shock profile, can develop a very steep front that causes oscillations in the box scheme.

In Chapter 5 we apply the mathematical and numerical analysis from Chapters 2 and 3 to the Flushing-through Model. We show that the box scheme is very robust in handling the different retardation speeds that can arise. A modified equation analysis is used to make predictions about how the different chemical pollutants behave. It is able to predict that the concentrations a and c can travel at different speeds for certain parameter values. The asymptotic analysis carried out to obtain a simpler system, also derived for the Linear Model in Chapter 2, again highlights the presence of diffusion but is shown to be very restrictive unless all the reactions are stiff. This leads us to conclude that for larger systems with varying speeds, an asymptotic expansion of the solution gives a poor approximation and so a numerical method must be used.

Finally, in Chapter 6 the box scheme is applied to model problems in 2D. Since only non-constant velocity fields are realistic we have extra complications in the numerical method. For non-smooth data the solution has more severe oscillations than for problems with constant coefficients. Guided by an understanding of 1D problems with variable speeds, we use a modified equation analysis to choose the spatial step for the numerical method (dependent on the time-step and the variable coefficient). This is a new technique and greatly reduces the oscillations. We also predict the best parameter to take for the time-weighting parameter (dependent on the time-step) which decreases them further. This can give a large improvement over the application of the box scheme with a constant spatial step which we will illustrate with numerical results.

Chapter 2

The Linear Model problem and extensions

The aim of this Chapter is to analyse the basic Linear Model and related models which describe the transport of chemical pollutants in groundwater flow. We use as an example the chromatography of a single solute: chromatography is a process by which a separation of chemical pollutants is obtained by selective adsorption on a solid medium (which we assume to be a rock). However, some of the techniques described can be readily applied to a more general problem, and so we present these techniques in a wider context.

We start with a statement of the problem and describe the standard initial and boundary conditions that will be used throughout the Thesis. Then, crude upper and lower bounds of the solution will be derived. In Section 2.2 we write down the exact solution of the Linear Model which is found using Laplace transforms; and has been derived in great detail in (Rhee et al. 1986). This gives us a means by which we can test the accuracy of the numerical solution described in Chapter 3 and, provided the parameters λ and μ are assumed large, we can approximate the exact solution to depict how the model behaves in practical situations. We deduce that, in this special case, the chemical pollutant moves at a slower speed (which we refer to as the reduced speed) than the advected speed. This is an important feature of these types of systems and will be seen throughout the Thesis.

Eventually we will be considering systems with various chemical pollutants and so it is likely that the parameters in the reaction terms will not be of similar orders of magnitude. This means we need to understand how the solution behaves in the Linear Model for a large range of λ and μ. In Section 2.3 we describe how the solution propagates for both large and small λ and μ and discuss the different phenomena for each. We then show how chemical engineers would obtain a simplified system by assuming that

the concentrations are in equilibrium (i.e. that the chemical reaction rates are fast compared with the rates of transport and so very quickly tend to a constant state). This is described in Section 2.4 and is referred to as the Equilibrium model; it gives the reduced speed that we found by approximating the Laplace transform solution. In Section 2.5 we go on to describe a method (using asymptotic analysis) which derives the Equilibrium model for general hyperbolic systems with relaxation where the source terms quickly tend to their equilibrium values. For the Linear Model this is equivalent to saying that λ and μ are large. A correction to the Equilibrium model can then be found for this system which we apply to the Linear Model. We call this the Improved-equilibrium model and it shows how diffusion is an important feature which is not obvious from the original formulation. We observe that this correction can approximate the solution very accurately in certain situations.

Lighthill & Whitham (1955) considered a linearised model to describe flood movement in long rivers and looked at the solution in transformed variables. This is obtained by applying a Heaviside transformation (very similar to a Laplace transformation) and they are able to show that the dynamic wave-front decays exponentially. Then an approximate form of the solution for large time was found. Lighthill & Whitham's (1955) model is a special case of the Linear Model so we cannot use the results directly; but, in Section 2.6, the same procedure is applied to the Linear Model to deduce similar behaviour. We again observe from this analysis that the maximum part of the solution wave travels downstream at the reduced speed that we have discussed above.

In these first few Sections we observe some interesting phenomena for large λ and μ. The exact solution found by Laplace transforms is valid for all values of these parameters but we can only make any observations about how the solution progresses over time by assuming they are large. The asymptotic analysis in Section 5 also requires this assumption. As stated above we would like to be able to make observations about the behaviour of these types of models for a large range of values of the parameters involved. We aim to prove that bounds exist on the speed of propagation for a general two equation model assuming certain lower bounds exist on the partial derivatives of the reaction term. In order to do this we must use the fact that the concentration a decays exponentially and so Section 2.7 concentrates on proving this result. Finally, in Section 2.8 the model is integrated over a given domain to obtain an expression relating the integrals on the boundaries. If the domain is extended to infinity we can use the result from Section 2.7 to prove that the concentration of chemical pollutants will move at a slower speed than the speed of advection for a general reaction term.

2.1 Problem statement and solution bounds

In this Chapter we consider the Linear Model

$$a_t + Va_x = -\lambda a + \mu b \tag{2.1}$$

$$b_t = \lambda a - \mu b, \tag{2.2}$$

where $\lambda \geq \mu > 0$ are constants. The standard initial and boundary conditions that we will use throughout this Thesis (unless otherwise stated) are

$$a(x,0) = b(x,0) = 0, \quad x \geq 0, \tag{2.3}$$

$$a(0,t) = g(t), \quad t > 0. \tag{2.4}$$

Note that only a needs to be prescribed on the $x = 0$ axis because (2.2) has no x-derivative for b and so can be integrated along $x = 0$ using (2.4). These initial and boundary conditions are typical of these problems: some chemical pollutant is placed at a point in space and we wish to observe how it spreads out over time at later points in space. From (2.3) it follows that

$$a_x(x,0) = 0, \tag{2.5}$$

and so we can use this and (2.3) in the model (2.1) and (2.2) to deduce that

$$a_t(x,0) = 0, \qquad b_t(x,0) = 0. \tag{2.6}$$

The character of the solution can be seen by integrating (2.1) and (2.2) along the two characteristics, $x = k_1$ and $x = Vt + k_2$ where k_1 and k_2 are arbitrary constants. Along these lines there are two ordinary differential equations to solve and we can express a in terms of b and vice versa. Hence

$$b(x,t) = \lambda \int_0^t a(x,r) e^{-\mu(t-r)} \mathrm{d}r, \tag{2.7}$$

and

$$a(x,t) = g\left(t - \tfrac{x}{V}\right) e^{-\frac{\lambda x}{V}} + \mu \int_{t-\frac{x}{V}}^t b\big(x - V(t-s), s\big) e^{-\lambda(t-s)} \mathrm{d}s. \tag{2.8}$$

In (2.7) we have exploited the fact that (2.2) only uses one independent variable t and so b can be written as an integral of a. We can immediately see from (2.7) that, if $a > 0$, it follows that $b > 0$. We now examine these integral expressions in more detail to obtain positivity results and upper bounds for both a and b. Suppose we set

$$p(r,t) = \frac{\mu e^{-\mu(t-r)}}{1 - e^{-\mu t}}, \tag{2.9}$$

and so

$$\int_0^t p(r,t)\,\mathrm{d}r = 1, \quad \forall\, t. \tag{2.10}$$

Then we can rewrite (2.7) as

$$b(x,t) = \frac{\lambda}{\mu}\left(1 - e^{-\mu t}\right) \int_0^t a(x,r)p(r,t)\,\mathrm{d}r. \tag{2.11}$$

Using (2.10) we can deduce the following:

$$\frac{\lambda}{\mu}\left(1 - e^{-\mu t}\right) \min_{0\le r\le t} a(x,r) \le b(x,t) \le \frac{\lambda}{\mu}\left(1 - e^{-\mu t}\right) \max_{0\le r\le t} a(x,r). \tag{2.12}$$

Let us now consider (2.8). We can eliminate b in the integral term by substituting the solution of b from (2.7). So

$$a(x,t) = g\left(t - \tfrac{x}{V}\right) e^{-\frac{\lambda x}{V}} + \mu\lambda \int_{t-\frac{x}{V}}^t e^{-\lambda(t-s)} \int_0^s a(x - V(t-s), r)e^{-\mu(s-r)}\,\mathrm{d}r\,\mathrm{d}s. \tag{2.13}$$

Suppose we now set

$$\bar{a}(x,t) = \int_0^t a(x,r)p(r,t)\,\mathrm{d}r, \tag{2.14}$$

where $p(\cdot)$ is defined by (2.9). Then from (2.11) we have

$$b(x,t) = \frac{\lambda}{\mu}\left(1 - e^{-\mu t}\right)\bar{a}(x,t). \tag{2.15}$$

Also set

$$q(s,t) = \frac{\lambda e^{-\lambda(t-s)}}{1 - \rho(x)}, \tag{2.16}$$

where

$$\rho(x) = e^{-\frac{\lambda x}{V}}, \tag{2.17}$$

so that

$$\int_{t-\frac{x}{V}}^t q(s,t)\,\mathrm{d}s = 1, \quad \forall\, t.$$

Then (2.13) becomes

$$a(x,t) = g\left(t - \tfrac{x}{V}\right)\rho(x) + \left[1 - \rho(x)\right] \int_{t-\frac{x}{V}}^t \bar{a}(x - V(t-s), s)\left(1 - e^{-\mu s}\right)q(s,t)\,\mathrm{d}s. \tag{2.18}$$

Equations (2.15) and (2.18) are the basis for our bounds on a and b:

1. If $g\left(t-\frac{x}{V}\right) \geq 0$ for $0 \leq t \leq T$ then

$$a(x,t) \geq 0, \qquad b(x,t) \geq 0 \quad \forall\ (x,t). \tag{2.19}$$

The reason for this is as follows: since $t - \frac{1}{V}x \leq s \leq t$ we have $t - s \geq 0$ and so $x - V(t-s) \leq x$. Thus the expression in the integral in (2.18) is positive as $\bar{a}(x - V(t-s), s)$ occurs at a point in space which is upwind. If $s = t - \frac{1}{V}x$ then $x - V(t-s) = 0$ which is on the $x = 0$ boundary where we know a is positive.

2. If $g\left(t-\frac{x}{V}\right) \leq A$ for $0 \leq t \leq T$ then $\forall\ (x,t)$ we have

$$a(x,t) \leq A \tag{2.20}$$
$$b(x,t) \leq \frac{\lambda}{\mu}\left(1 - e^{-\mu t}\right)A, \tag{2.21}$$

which come from (2.18) and (2.11) respectively.

2.2 The exact Laplace Transform solution

The exact solution of (2.1) and (2.2) with conditions (2.3) and (2.4) can be found using Laplace transforms. The details of this transformation are rather technical and so we refer the reader to (Rhee et al. 1986, pages 145–159). The general solution is given by

$$a(x,t) = \begin{cases} 0, & \text{if } t \leq \frac{x}{V} \\ g\left(t-\frac{x}{V}\right)e^{-\frac{\lambda x}{V}} + \mu\int_0^{t-\frac{x}{V}} g\left(t-\frac{x}{V}-s\right)e^{-\frac{\lambda x}{V}-\mu s}G(s)\,\mathrm{d}s, & \text{if } t > \frac{x}{V}, \end{cases} \tag{2.22}$$

where

$$G(s) = \sqrt{\frac{\lambda x}{\mu V s}}\, I_1\left(2\sqrt{\frac{\lambda\mu x s}{V}}\right), \tag{2.23}$$

and $I_1(\cdot)$ is the modified Bessel function of the first kind (Abramowitz & Stegun 1965). Once the solution of a has been found we can write down an expression for b by substituting (2.22) into (2.7). Later in this Chapter we will show that the Linear Model can be written as a second order equation in only one variable. In Appendix A (Section A.3) we derive (2.22) by integrating this second order equation over a region bounded by the two characteristics. This is a much easier way to find the solution than using Laplace Transforms.

Also in Appendix A (Section A.1) we derive a simplification to the solution a given in (2.22) when the argument of the Bessel function is large. Consider the situation where $g(t)$ is an injection of a short pulse of chemical pollutants into the groundwater at $x = 0$, and so

$$g(t) = 0, \quad t > \delta, \qquad \int_0^{\delta} g(t)\,\mathrm{d}t = \alpha, \tag{2.24}$$

where $\alpha > 0$ is constant. After some manipulation of (2.22) we find that

$$a(x,t) \sim \frac{\alpha\mu}{2\sqrt{\pi}} \left[\frac{\lambda x/V}{(\mu(t-x/V))^3} \right]^{1/4} \exp\left[-\left(\sqrt{\mu\left(t-\tfrac{x}{V}\right)} - \sqrt{\tfrac{\lambda x}{V}} \right)^2 \right], \tag{2.25}$$

which holds for $t > \frac{x}{V}$ ($a = 0$ for $t \leq \frac{x}{V}$). If we fix x then the maximum value (with respect to t) of the solution $a(x,t)$ occurs when the exponential term is zero, i.e.

$$t = \frac{\lambda+\mu}{\mu V} x = \frac{1}{V'} x, \tag{2.26}$$

where

$$V' = \frac{V\mu}{\lambda+\mu}. \tag{2.27}$$

Note that because λ and μ are assumed to be positive we have $V' < V$. We call V' the *reduced (or retarded) speed.* Hence, when λ and μ are large, the peak moves with the reduced speed. This is a critical phenomenon of the model: there is a delay in output of the chemical pollutants due to the chemical interaction. It is this speed that makes the Linear Model, which on first appearance seems quite simple, a challenge to analyse. It is known that systems of this form often exhibit retardation, i.e. the reaction term produces a speed which is slower than the speed of the groundwater itself. In fact, we can observe this directly by adding together equations (2.1) and (2.2). This gives

$$(a+b)_t + Va_x = 0, \tag{2.28}$$

which is of the form $u_t + f_x = 0$ where $u = a + b$ and $f = Va$. If we integrate this over a rectangular region with steady left and right states we deduce a steady speed $[Va]/[a+b]$, where $[\cdot]$ denotes the jump. This is simply the Rankine-Hugoniot jump condition for scalar problems (LeVeque 1992, page 31). Hence we can expect a speed less than V.

2.3 Large λ and μ versus small λ and μ

In Figures 2-1 and 2-2 we plot the exact solution a against t, given in (2.22), for various λ and μ at three stages of x. The boundary condition is a square pulse, as shown to the left of the vertical line in the top plots. In Figure 2-1 the parameters are small and in the left three plots the solution moves at the advection speed V, as expected. However, as λ is increased (the right plots), the solution now moves much more slowly. It is not quite travelling at the reduced speed (which is $V' = \frac{1}{4}$ in this case) but is closer than when $\lambda = \mu = 1$. In Figure 2-2 we have increased λ and μ further: the pulse is now moving much closer to the reduced speed V' (which equals 1/10). These Figures agree with the analysis above: for large λ and μ the pulse moves at the reduced speed.

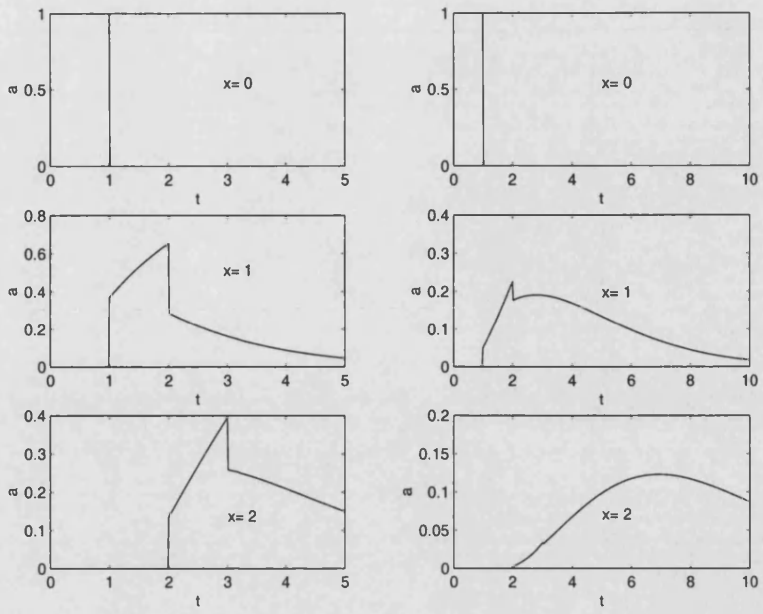

Figure 2-1: *The Laplace transform solution* $a(x,t)$ *at* $x = 0$ *(top plots),* $x = 1$ *(middle plots) and* $x = 2$ *(bottom plots) for* $V = 1$. *In the left plots* $\lambda = \mu = 1$ *and in the right plots* $\lambda = 3$, $\mu = 1$.

We can also use these plots to show how quickly the solution is smoothed, even for small λ and μ. Suppose we consider the Linear Model in the form of (2.1) and (2.2) and assume that λt and μt are small. In these early stages the source term is not dominant and a is advected with speed V (being fed by the boundary data and the μb term, and damped by $-\lambda a$); while b is not advected at all, but merely fed by the λa term and damped by $-\mu b$. Hence the solution for a will behave like the linear advection equation with speed V. In the middle plots in Figure 2-1 the solution switches on at $t = 1$ and builds up until $t = 2$. This can be explained using (2.8). At the switch-on we can ignore the integral term and so, for $\lambda = \mu = 1$, $a = g(0)e^{-1} \approx 0.369$ which is the value given in the plot. As time progresses the integral term is added and so the pulse increases to a maximum at $t = 2$. When $\lambda = 3$ and $\mu = 1$ the switch-on is now much smaller at $t = 1$ ($a = g(0)e^{-3} \approx 0.0498$) and builds up through b to a much smaller value than when $\lambda = 1$. Thus, even for λ and μ relatively small, there is diffusion in the model; the pulse starts sharp but very quickly reduces due to the negative exponential term in (2.8). We will see this in Section 2.5 when an asymptotic analysis is performed.

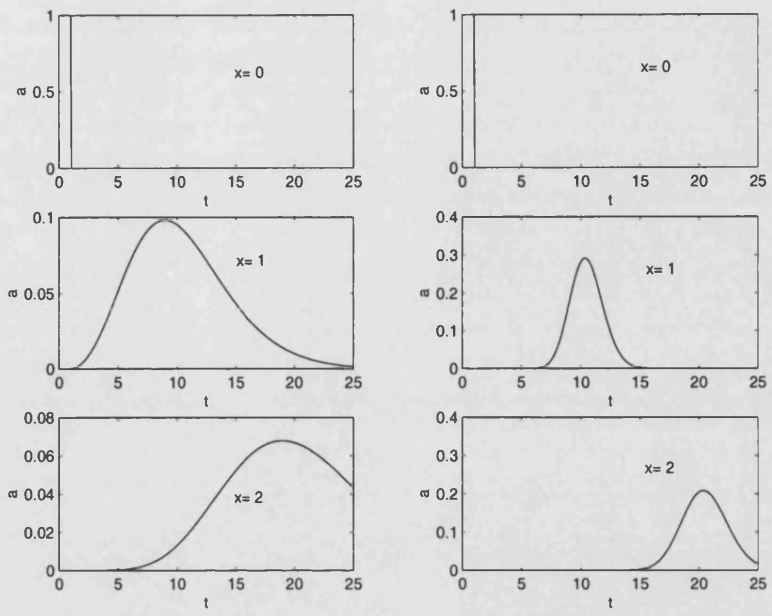

Figure 2-2: *The Laplace transform solution $a(x,t)$ at $x = 0$ (top plots), $x = 1$ (middle plots) and $x = 2$ (bottom plots) for $V = 1$. In the left plots $\lambda = 9$, $\mu = 1$ and in the right plots $\lambda = 90$, $\mu = 10$.*

2.4 The Equilibrium model

For some systems the chemical reaction rates are so fast compared to the rates of transport that the chemical reactions can be modelled by assuming that they proceed instantaneously to equilibrium. This leads to a solution procedure which gives us an *Equilibrium model.* Suppose we add (2.1) and (2.2) to give the model in the form

$$a_t + b_t + V a_x = 0 \tag{2.29}$$

$$b_t = \lambda a - \mu b. \tag{2.30}$$

The first equation is simply a conservation law and can be written as

$$c_t + V a_x = 0, \tag{2.31}$$

where $c := a + b$ is the total concentration of chemical pollutants. We now assume that $\lambda > \mu \gg 1$. If (2.30) is divided by λ then

$$\frac{1}{\lambda} b_t = a - \frac{\mu}{\lambda} b.$$

Neglecting the left hand side which is small (due the chemical reactions immediately becoming non time-dependent) means we can express b as a function of a leading to

the equilibrium condition

$$b = \frac{\lambda}{\mu} a. \tag{2.32}$$

Hence the concentration b of chemical pollutant in the rock is proportional to the solute concentration a (as discussed in Section 1.3 of the Introduction). We can substitute (2.32) into (2.29) to give the following linear advection equation for a:

$$a_t + \frac{V\mu}{\lambda + \mu} a_x = 0. \tag{2.33}$$

This is our Equilibrium model which moves at the reduced speed found in the analysis of the exact Laplace transform solution when λ and μ were assumed to be large. The exact solution of (2.33) is simply

$$a(x,t) = g\left(t - \frac{1}{V'}x\right). \tag{2.34}$$

and so the boundary condition is propagated with speed V' without changing shape. Thus the Equilibrium model gives the correct speed for the solution when λ and μ are large. However, as we shall see from comparison with the exact Laplace transform solution in Section 2.5.3, the height and shape of the pulse are very inaccurate. It is obvious that (2.33) gives the wrong speed when λ and μ are small. The left plots in Figure 2-1 consider $\lambda = \mu = 1$ and we observe that the solution of a moves with speed V $(= 1)$. However, the Equilibrium model predicts a reduced speed of $\frac{1}{2}$ and so is very inaccurate. The Equilibrium model will only ever give a good approximation when λ and μ are both large and approximately equal (see Figure 2-3).

Note that the same conservation law holds for the total concentration $c := a + b$. From the equilibrium condition (2.32) we have

$$a = c - b = c - \frac{\lambda}{\mu} a, \tag{2.35}$$

and so

$$a = \frac{\mu}{\lambda + \mu} c.$$

This can be substituted into the conservation law (2.31) to give the equilibrium model

$$c_t + V' c_x = 0. \tag{2.36}$$

2.5 The Improved-equilibrium model

In this Section we derive a better approximation to the Linear Model (2.29) and (2.30) which is a correction to the Equilibrium Model and more accurately describes the characteristics of the solution for large λ and μ. The method, as described in (Chen, Levermore & Liu 1994), can be applied to a general hyperbolic system with relaxation, i.e. a system of hyperbolic conservation laws with source terms whose effect is to create a time difference between the various concentrations in the system. This is defined as a relaxation effect and the time difference (or lag) is referred to as the relaxation time. The effect of relaxation is basically to study how the concentrations relax to their equilibrium values. We will describe the general method (from now on referred to as the *Improved-equilibrium model*) and then apply it to the Linear Model.

It should be noted that (Liu 1987) also describes a similar method for hyperbolic conservation laws with relaxation effects which leads to the same simplified system as in (Chen et al. 1994). This differs from the analysis in (Chen et al. 1994) because it only considers a pair of quasilinear hyperbolic equations, a conservation law coupled with a rate equation. A time-asymptotic expansion is used to derive a simplified system.

2.5.1 The general case

Consider a general hyperbolic system with relaxation in the form

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}) + \frac{1}{\epsilon}\mathbf{S}(\mathbf{u}) = \mathbf{0}, \tag{2.37}$$

where the state vector $\mathbf{u}$ of dimension p belongs to some given subset Ω of $\mathbb{R}^p$ and the flux function $\mathbf{f}$ is such that, for any $\mathbf{u} \in \Omega$, the matrix $\partial \mathbf{f}/\partial \mathbf{u}$ has real eigenvalues. The quantity $1/\epsilon$ is the relaxation time. The first order system

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}) = \mathbf{0}, \tag{2.38}$$

is a system of hyperbolic conservation laws.

Using the analysis of (Bereux & Sainsaulieu 1997) and (Chen et al. 1994), two approximations of this system are now derived. In the sense of (Whitham 1974) and (Liu 1987) the source terms are relaxation terms if there exists a constant $r \times p$ matrix Q with rank $r < p$ such that

$$Q\mathbf{S}(\mathbf{u}) = \mathbf{0}, \qquad \forall\, \mathbf{u} \in \Omega, \tag{2.39}$$

and if, for any given $\mathbf{u}^0 \in \Omega$, the differential equation

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = -\frac{1}{\epsilon}\mathbf{S}(\mathbf{u}), \qquad \mathbf{u}(0) = \mathbf{u}^0, \tag{2.40}$$

defines a function $\mathbf{u} : \mathbb{R}^+ \to \Omega$ such that

$$\lim_{t \to +\infty} \mathbf{u}(t) = Q\mathbf{u}^0.$$

Define $\mathcal{M}$ to be the equilibrium manifold or the manifold of local equilibria for $\mathbf{S}$

$$\mathcal{M} := \{\mathbf{u} \in \Omega \subset \mathbb{R}^p \mid \mathbf{S}(\mathbf{u}) = \mathbf{0}\}. \tag{2.41}$$

Moreover, define the set ω to be given by

$$\omega := \{\mathbf{c} \in \mathbb{R}^r \mid \mathbf{c} = Q\mathbf{u},\ \mathbf{u} \in \Omega\}. \tag{2.42}$$

Then assume there exists a mapping $\mathcal{E} : \omega \to \mathcal{M}$, so that

$$\mathbf{c} \in \omega \implies \mathbf{S}(\mathcal{E}(\mathbf{c})) = \mathbf{0}, \tag{2.43}$$

and also

$$Q\mathcal{E}(\mathbf{c}) = \mathbf{c}, \quad \forall\, \mathbf{c} \in \omega. \tag{2.44}$$

Assume further that $Q : \mathcal{M} \to \omega$ defines a bijection. Then if $\mathbf{u} \in \Omega$ is such that $\mathbf{S}(\mathbf{u}) = \mathbf{0}$ it follows that $\mathbf{u} = \mathcal{E}(Q\mathbf{u})$. If $\mathbf{u}$ is a solution of (2.37) then since Q is a constant matrix and $Q\mathbf{S}(\mathbf{u}) = \mathbf{0}$, $\forall\, \mathbf{u} \in \Omega$, we have

$$\frac{\partial Q\mathbf{u}}{\partial t} + \frac{\partial}{\partial x} Q\mathbf{f}(\mathbf{u}) = \mathbf{0}. \tag{2.45}$$

When ϵ is small we expect that the solution $\mathbf{u}$ will be close to the equilibrium manifold $\mathcal{M}$ (since $\mathbf{S}(\mathbf{u}) = \mathbf{0}$ then dominates (2.37)) and, more precisely,

$$\mathbf{u} = \mathcal{E}(\mathbf{c}) + O(\epsilon),$$

where $\mathbf{c} = Q\mathbf{u}$. At zeroth order in ϵ it is deduced that $\mathbf{c}$ is a solution of

$$\frac{\partial \mathbf{c}}{\partial t} + \frac{\partial}{\partial x} \mathbf{g}(\mathbf{c}) = \mathbf{0}, \tag{2.46}$$

where the flux function $\mathbf{g}$ is given by

$$\mathbf{g}(\mathbf{c}) = Q\mathbf{f}(\mathcal{E}(\mathbf{c})). \tag{2.47}$$

This is known as the equilibrium equation. Since the image of $\mathcal{E}$ is just the equilibrium manifold, (2.47) can be written as

$$\mathbf{g}(\mathbf{c}) = Q\mathbf{f}(\mathcal{M}[\mathbf{c}]). \tag{2.48}$$

Following the method given in (Chen et al. 1994) a first order expansion can be obtained. Suppose that the dynamics of $\mathbf{u}$ are governed by the system (2.46). The evolution of $\mathbf{u} = \mathcal{M}[\mathbf{c}]$ is then given by

$$\frac{\partial \mathbf{u}}{\partial t} = \frac{\partial}{\partial \mathbf{c}}(\mathcal{M}[\mathbf{c}])\frac{\partial \mathbf{c}}{\partial t} = -\frac{\partial}{\partial \mathbf{c}}(\mathcal{M}[\mathbf{c}])\frac{\partial}{\partial x}Q\mathbf{f}(\mathcal{M}[\mathbf{c}]),$$

(using (2.46) and (2.48) to eliminate $\partial \mathbf{c}/\partial t$). Hence (2.37) becomes

$$\begin{aligned}
\mathbf{0} &= \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}) + \frac{1}{\epsilon}\mathbf{S}(\mathbf{u}) \\
&= -\frac{\partial}{\partial \mathbf{c}}(\mathcal{M}[\mathbf{c}])\frac{\partial}{\partial x}Q\mathbf{f}(\mathcal{M}[\mathbf{c}]) + \frac{\partial}{\partial x}\mathbf{f}(\mathcal{M}[\mathbf{c}]) + \frac{1}{\epsilon}\mathbf{S}(\mathcal{M}[\mathbf{c}]) \\
&= \left(I - \frac{\partial}{\partial \mathbf{c}}(\mathcal{M}[\mathbf{c}])Q\right)\frac{\partial}{\partial x}\mathbf{f}(\mathcal{M}[\mathbf{c}]) + \frac{1}{\epsilon}\mathbf{S}(\mathcal{M}[\mathbf{c}]),
\end{aligned}$$

assuming Q is independent of x and I is the identity matrix.

We aim to find $\mathcal{M}$ such that the right hand side of the above expression is small. Set

$$\mathcal{M}^\epsilon[\mathbf{c}] = \mathcal{E}(\mathbf{c}) + \epsilon\mathcal{M}^{(1)}[\mathbf{c}] + \epsilon^2\mathcal{M}^{(2)}[\mathbf{c}] + \ldots. \tag{2.49}$$

Then $\mathcal{M}^\epsilon[\mathbf{c}]$ has to satisfy the following two conditions:

$$\left(I - \frac{\partial}{\partial \mathbf{c}}(\mathcal{M}^\epsilon[\mathbf{c}])Q\right)\frac{\partial}{\partial x}\mathbf{f}(\mathcal{M}^\epsilon[\mathbf{c}]) + \frac{1}{\epsilon}\mathbf{S}(\mathcal{M}^\epsilon[\mathbf{c}]) = \mathbf{0} \tag{2.50}$$

$$Q\mathcal{M}^\epsilon[\mathbf{c}] = \mathbf{c}. \tag{2.51}$$

The second comes from the fact that $Q\mathcal{E}(\mathbf{c}) = \mathbf{c}$ and $\mathcal{E}(\mathbf{c}) = \mathcal{M}[\mathbf{c}]$ in the equilibrium approximation. Substituting the expansion (2.49) into (2.50) and (2.51) and matching the constant terms gives

$$\left(I - \frac{\partial}{\partial \mathbf{c}}(\mathcal{E}(\mathbf{c}))Q\right)\frac{\partial}{\partial x}\mathbf{f}(\mathcal{E}(\mathbf{c})) + \frac{\partial}{\partial \mathbf{u}}\mathbf{S}(\mathcal{E}(c))\mathcal{M}^{(1)}[\mathbf{c}] = \mathbf{0} \tag{2.52}$$

$$Q\mathcal{E}(\mathbf{c}) + \epsilon Q\mathcal{M}^{(1)}[\mathbf{c}] = \mathbf{c}, \tag{2.53}$$

using the fact that

$$\mathbf{S}(\mathcal{M}^\epsilon[\mathbf{c}]) = \mathbf{S}(\mathcal{E}(\mathbf{c})) + \epsilon\frac{\partial}{\partial \mathbf{u}}\mathbf{S}(\mathcal{E}(\mathbf{c}))\mathcal{M}^{(1)}[\mathbf{c}] = \epsilon\frac{\partial}{\partial \mathbf{u}}\mathbf{S}(\mathcal{E}(\mathbf{c}))\mathcal{M}^{(1)}[\mathbf{c}],$$

(and similarly for $\mathbf{f}(\mathcal{M}^\epsilon[\mathbf{c}])$) up to and including the ϵ term, and $\mathbf{S}(\mathcal{E}(\mathbf{c})) = \mathbf{0}$. Since $Q\mathcal{E}(\mathbf{c}) = \mathbf{c}$ the expression given in (2.53) reduces to

$$Q\mathcal{M}^{(1)}[\mathbf{c}] = \mathbf{0}. \tag{2.54}$$

If ϵ is assumed to be small then $\mathcal{M}^{\epsilon}[\mathbf{c}]$ is taken to be just the first two terms in (2.49), where $\mathcal{M}^{(1)}[\mathbf{c}]$ is found by solving (2.52) with (2.54). The matrix $\frac{\partial}{\partial \mathbf{u}}\mathbf{S}(\mathcal{E}(c))$ is singular and so, for the general case, $\mathcal{M}^{(1)}[\mathbf{c}]$ cannot be written down directly.

Finally if $\mathcal{M}^{\epsilon}[\mathbf{c}]$ is substituted for $\mathcal{M}[\mathbf{c}]$ in (2.46) and (2.48) then the correction to the local equilibrium approximation is obtained

$$\frac{\partial \mathbf{c}}{\partial t} + \frac{\partial}{\partial x} Q\mathbf{f}\left(\mathcal{E}(\mathbf{c}) + \epsilon \mathcal{M}^{(1)}[\mathbf{c}]\right) = \mathbf{0}. \tag{2.55}$$

This technique is now applied to the Linear Model.

2.5.2 The Linear Model

The Linear Model defined by (2.29) and (2.30) can be written in the form of (2.37) where

$$\mathbf{u} = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \mathbf{f}(\mathbf{u}) = \begin{bmatrix} Va \\ 0 \end{bmatrix}, \quad \mathbf{S}(\mathbf{u}) = \begin{bmatrix} a - \frac{\mu}{\lambda} b \\ -a + \frac{\mu}{\lambda} b \end{bmatrix}, \tag{2.56}$$

with $\epsilon = \frac{1}{\lambda}$. If we choose

$$Q = [1,\ 1], \quad \mathcal{E}(c) = \begin{bmatrix} \frac{\mu}{\lambda+\mu} c \\ \frac{\lambda}{\lambda+\mu} c \end{bmatrix},$$

then the conditions given in (2.39), (2.43) and (2.44) are satisfied and $c := Q\mathbf{u} = a + b$ is the total concentration. Since $Q\mathbf{f}(\mathcal{E}(c)) = \frac{V\mu}{\lambda+\mu} c$, the equilibrium model given by (2.46) and (2.47) becomes

$$c_t + \frac{V\mu}{\lambda + \mu} c_x = 0. \tag{2.57}$$

This is the same equation that was derived in Section 2.4. $\mathcal{M}^{(1)}[c]$ can now be found using (2.52) and (2.53). For the linear model the first term of (2.52) is simplified to

$$\left(I - \frac{\partial}{\partial c}(\mathcal{E}(c))Q\right) \frac{\partial}{\partial x}\mathbf{f}(\mathcal{E}(c)) = \begin{bmatrix} \frac{V\lambda\mu}{(\lambda+\mu)^2} c_x \\ -\frac{V\lambda\mu}{(\lambda+\mu)^2} c_x \end{bmatrix}.$$

Solving the system (2.52) with $Q\mathcal{M}^{(1)}[c] = 0$ leads to the solution

$$\mathcal{M}^{(1)}[c] = \begin{bmatrix} -\frac{V\lambda^2\mu}{(\lambda+\mu)^3} c_x \\ \frac{V\lambda^2\mu}{(\lambda+\mu)^3} c_x \end{bmatrix}.$$

Then, from (2.49), $\mathcal{M}^{\epsilon}[c] = \mathcal{E}(c) + \epsilon \mathcal{M}^{(1)}[c]$ (up to and including the ϵ term) and this replaces $\mathcal{M}[c]$ in (2.48). Substituting this back into (2.46) gives the correction to the

local equilibrium approximation, which we call the *Improved-equilibrium model*

$$c_t + \frac{V\mu}{\lambda+\mu}c_x - \frac{V^2\lambda\mu}{(\lambda+\mu)^3}c_{xx} = 0. \tag{2.58}$$

This introduces a diffusion term: we know that there is diffusion in the model (from observing plots of the exact Laplace transform solution in Figures 2-1 and 2-2) even though there are no diffusion terms in the model equations.

2.5.3 Discussion of the Improved-equilibrium model

The equation (2.58) gives a better approximation than the equilibrium model because it includes a diffusion term. However it can only give a good approximation to the Linear Model when λ and μ are large. We are interested in a large range of values for these parameters. In the quarter plane $x > 0$, $t > 0$, the exact solution of (2.58), with the conditions $c(x,0) = 0$ and $c(0,t) = h(t)$, can be found. Here we just quote the result (the details are given in (DuChateau & Zachmann 1989, page 239)):

$$c(x,t) = \int_0^t h(t-\tau)\frac{x}{\sqrt{4\pi D\tau^3}}\exp\left[-\frac{1}{4D\tau}(x-V'\tau)^2\right]\,\mathrm{d}\tau, \tag{2.59}$$

where

$$V' = \frac{V\mu}{\lambda+\mu}, \qquad D = \frac{V^2\lambda\mu}{(\lambda+\mu)^3}. \tag{2.60}$$

In Figure 2-3 we show plots of a for the exact Laplace transform solution (solid line), the equilibrium model (dashed line) and the Improved-equilibrium model ($\times$'s) for four different values of λ and μ (note that $V = 1$ in all these cases). The dotted line indicates the boundary condition which is a square pulse. These plots illustrate that, as λ and μ get large, the Improved-equilibrium model approximates the exact solution very closely and the equilibrium model starts to move at the correct speed. Hence the Improved-equilibrium model can give an accurate representation of the solution when λ and μ are large. Also, it clearly shows the diffusion in the model.

2.6 Lighthill-Whitham analysis

2.6.1 Introduction

Lighthill & Whitham (1955) wrote a paper about the theory of a distinctive type of wave motion, which arises in any one-dimensional flow problem. This class of wave motions is physically quite distinct from the classical wave motions encountered in dynamical systems. They exist if, to a sufficient approximation, there is a functional relationship between the flow, the concentration and the position. On this assumption the wave property follows from the equation of continuity alone. The waves are described as

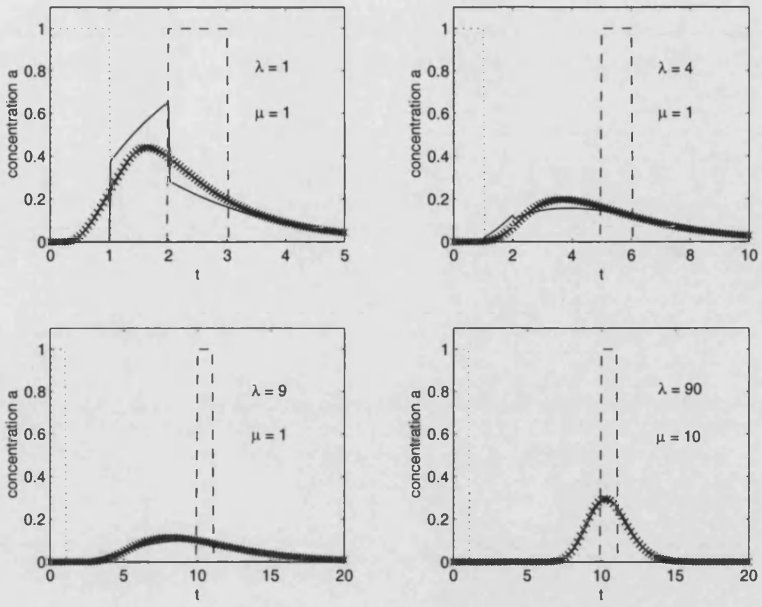

Figure 2-3: *Comparison of the exact Laplace transform solution (a solid line), the equilibrium model (dashed line) and the Improved-equilibrium model (×'s). The dotted line shows the boundary condition and in all cases* $V = 1$. *Top left:* $\lambda = \mu = 1$ *(and so* $V' = 1/2$*), top right:* $\lambda = 4$ *and* $\mu = 1$ *(and so* $V' = 1/5$*), bottom left:* $\lambda = 9$ *and* $\mu = 1$ *(and so* $V' = 1/10$*) and bottom right:* $\lambda = 90$ *and* $\mu = 10$ *(and so* $V' = 1/10$*).*

"kinematic". In contrast, the classical wave motions would be described as "dynamic" waves, depending on Newton's second law of motion. An important difference is that kinematic waves possess only one wave velocity at each point, whereas dynamic waves possess at least two. Kinematic waves are not dispersive, but they suffer change of form due to non-linearity. In Section 3 of (Lighthill & Whitham 1955) it is shown that the mathematical relations between kinematic and dynamic waves exist by deriving a model which describes flood movement in long rivers. Define h to be the "stage" at each point on the river, i.e. the height of the free surface above a certain reference plane, and v to be the mean velocity. Attention is restricted to the linear theory of small disturbances. This linear approximation is quite severe but a complete solution containing both *kinematic and dynamic waves* can be obtained. The model is given by

$$u_t + v_0 u_x + g\eta_x + gS_0\left(\frac{2u}{v_0} - \frac{\eta}{h_0}\right) = 0 \tag{2.61}$$

$$\eta_t + v_0\eta_x + h_0 u_x = 0, \tag{2.62}$$

where $v = v_0 + u$ and $h = h_0 + \eta$ have been substituted into the nonlinear model and only first-order terms in u and η have been retained. Note that S_0 is the slope of the reference surface and g is the usual acceleration due to gravity. This model has four

parameters: v_0, h_0, S_0 and g.

We would like to be able to derive the Linear Model from this pair of equations because (2.62) is a conservation equation and (2.61) is a reaction equation. The Linear Model only has three parameters (V, λ and μ) but, for a special case, we should be able to write the Linear Model in the above form. For flood waves the local term is the bed friction whereas for the Linear Model it is the chemical reaction. The balance between the bed friction and the x derivative dominates and this will give a different speed which is exactly analogous to the reduced speed. We will apply this analysis to the Linear Model to obtain an approximate form of the solution for large time and will be able to deduce the reduced speed.

Suppose we multiply (2.61) by v_0 and (2.62) by g and subtract the resulting equations. This eliminates the η_x term from (2.61) to give

$$v_0 u_t + g\eta_t + (v_0^2 - gh_0)u_x + v_0 g S_0 \left(\frac{2u}{v_0} - \frac{\eta}{h_0}\right) = 0. \tag{2.63}$$

Let us now consider the Linear Model written in terms of c and b, i.e.

$$c_t + V(c_x - b_x) = 0 \tag{2.64}$$

$$b_t = \lambda c - (\lambda + \mu)b. \tag{2.65}$$

On comparing (2.63) and (2.65), it is clear that the x derivative in (2.63) must be eliminated by setting $v_0^2 = gh_0$. Then we have eliminated one of their parameters (say h_0). Hence, Lighthill & Whitham's (1955) model is now (with (2.62) multiplied by g)

$$g\eta_t + v_0 g\eta_x + v_0^2 u_x = 0. \tag{2.66}$$

$$v_0 u_t - g\eta_t + 2S_0 gu - \frac{v_0 g S_0}{h_0}\eta = 0 \tag{2.67}$$

Also, from comparing (2.65) and (2.67), it is clear that we must define $b := \beta(v_0 u - g\eta)$, where β is some constant to be determined. After some manipulation we can write (2.66) and (2.67) as

$$\eta_t + 2v_0\left(\eta_x + \frac{1}{2\beta g}b_x\right) = 0 \tag{2.68}$$

$$b_t = -\frac{\beta g^2 S_0}{v_0}\eta - \frac{2S_0 g}{v_0}b. \tag{2.69}$$

Then, if

$$c = \eta, \quad V = 2v_0, \quad \frac{1}{2\beta g} = -1, \tag{2.70}$$

equation (2.68) is simply (2.64); also (2.69) becomes

$$b_t = \frac{gS_0}{2v_0}c - \frac{2S_0}{v_0}b. \tag{2.71}$$

Finally, comparing this to (2.65) means we must have

$$\lambda = \frac{gS_0}{2v_0}, \quad \lambda + \mu = \frac{2S_0}{v_0}. \tag{2.72}$$

Hence $\mu = \frac{3gS_0}{2v_0} = 3\lambda$. So, Lighthill and Whitham's model, (2.61) and (2.62), is a special case of the Linear Model with $\mu = 3\lambda$. We therefore cannot utilise their results directly to convert from their notation into ours by substituting our definitions of u and η in terms of b and c. However, we can apply the same procedure.

2.6.2 The second order equation for the Linear Model

Lighthill & Whitham (1955) derive a single equation for η by differentiating (2.61) with respect to x and substituting for u_x from (2.62). This technique is also carried out in (Guenther & Lee 1988, page 212) for an identical system to the Linear Model except with an added diffusion term. We now derive a single second order equation in one variable for the Linear Model. Let us differentiate (2.64) with respect to t. Then

$$c_{tt} + V(c_{xt} - b_{xt}) = 0. \tag{2.73}$$

The b_{xt} term needs to be eliminated and so we differentiate (2.65) with respect to x to obtain

$$b_{xt} = \lambda c_x - (\lambda + \mu)b_x, \tag{2.74}$$

Now, from the conservation equation (2.64), $b_x = \frac{1}{V}(c_t + Vc_x)$. So, we obtain an expression for b_{xt} in terms of c using (2.74)

$$Vb_{xt} = -\mu V c_x - (\lambda + \mu)c_t. \tag{2.75}$$

Substituting this into (2.73) leads to

$$c_{tt} + Vc_{xt} + (\lambda + \mu)c_t + V\mu c_x = 0. \tag{2.76}$$

Note that the same equation holds for a (since the problem is linear).

We can write (2.76) in the form

$$\frac{1}{\lambda + \mu}\frac{\partial}{\partial t}\left[\frac{\partial}{\partial t} + V\frac{\partial}{\partial x}\right]c + \left[\frac{\partial}{\partial t} + \frac{V\mu}{\lambda + \mu}\frac{\partial}{\partial x}\right]c = 0. \tag{2.77}$$

Whitham (1974, pages 339-359) discusses the phenomenon of *wave hierarchies*: this is the situation when waves of different orders appear in the same problem. These can easily be seen by examining the factored operators in (2.77). If the lower order terms were absent (i.e. $\lambda + \mu \approx 0$) the general solution would be

$$c(x,t) = c_1(x) + c_2(x - Vt). \tag{2.78}$$

If the higher order terms were absent (i.e. $\lambda + \mu \approx \infty$) the general solution would be

$$c(x,t) = c^0\left(t - \frac{\lambda+\mu}{V\mu}x\right). \tag{2.79}$$

We now follow the procedure discussed in (Whitham 1974, pages 342-350) and consider the exact solution of this problem. Lighthill & Whitham (1955) also outline a similar method and we use both references as they contain different interpretations which are insightful for our model.

2.6.3 Heaviside calculus

First we specify the following conditions to ensure a well posed problem:

$$c(x,0) = c_t(x,0) = 0, \qquad x > 0, \tag{2.80}$$

$$c(0,t) = h(t), \qquad t > 0. \tag{2.81}$$

Note that $h(t) = a(0,t) + b(0,t)$, and $b(0,t)$ can easily be found from (2.7).

A Heaviside transformation (Jeffreys & Jeffreys 1972, page 212-213) is now applied which differs from the Laplace Transform by an additional factor p. So define

$$H(x,p) := p\int_0^\infty e^{-pt}c(x,t)\,\mathrm{d}t. \tag{2.82}$$

Then, setting

$$\eta = \frac{1}{\lambda+\mu}, \qquad V' = \frac{V\mu}{\lambda+\mu}, \tag{2.83}$$

and applying the integration in (2.82) to (2.77) leads to

$$[\eta V p + V']H_x + p[\eta p + 1]H = 0. \tag{2.84}$$

The general solution is

$$H(x,p) = A(p)e^{xP(p)}, \tag{2.85}$$

where

$$P(p) = -\frac{p(\eta p+1)}{\eta V p + V'}. \tag{2.86}$$

The function $A(p)$ is determined from the boundary condition (2.81). It is simply the Laplace transform of $h(t)$, i.e.

$$A(p) = p\int_0^\infty h(t)e^{-pt}\,\mathrm{d}t. \tag{2.87}$$

The interpretation of (2.85) (i.e. converting this back to $c(x,t)$ by applying the inverse Heaviside transform) gives an integral involving a Bessel function. This is to be expected since the exact solution, stated using Laplace transforms in Section 2.2, also involves a modified Bessel function of the first kind. The details are complicated and we can deduce the results required directly from (2.85).

The solution c can be expressed in terms of $H(x,p)$ by the contour integral

$$c(x,t) = \frac{1}{2\pi i}\int_{l-i\infty}^{l+i\infty}\frac{A(p)}{p}\exp\{pt + P(p)x\}\,\mathrm{d}p, \tag{2.88}$$

where l is so large that all the singularities of $H(x,p)$ lie to the left of the path of integration. The limits of this integral mean we are integrating over the set $\{p \in \mathbb{C} \mid \mathrm{Re}(p) = l\}$.

We are interested in the behaviour of the solution near the dynamic wave-front. The values of c near this wave-front correspond to the values of $H(x,p)$ for large p. Using the fact that $V' = V\mu\eta$, we can expand (2.86) to obtain an expression for $P(p)$ when p is large

$$\begin{aligned} P(p) &= -\left(\frac{p}{V}+\frac{1}{V\eta}\right)\left[1+\frac{\mu}{p}\right]^{-1} \\ &= -\frac{p}{V}-\frac{1}{V\eta}(1-\eta\mu)+O\left(\frac{1}{p}\right) \qquad (2.89) \\ &= -\frac{p}{V}-\frac{\lambda}{V}+O\left(\frac{1}{p}\right). \qquad (2.90) \end{aligned}$$

Substituting this expansion into (2.88) gives the approximation (see (Whitham 1974, page 344))

$$c \approx h\left(t-\frac{x}{V}\right)\exp\left\{-\frac{\lambda x}{V}\right\}. \tag{2.91}$$

Further terms in the series can be obtained by continuing the expansion of $e^{P(p)x}$ for large p. This expression is valid near the wavefront. It shows that the first disturbance propagates out with the V waves, but is damped exponentially and becomes negligible in a distance of order $V\eta$ (which can be seen from (2.89)). As $\eta \to 0$, this disturbance becomes negligible for all $x > 0$ which agrees with the reduced description from (2.79).

We now ask where the main disturbance described by (2.88) is found. To do this, the behaviour on the family of lines where x/t is constant is investigated, since each one of these represents the path of a wave moving with constant velocity. Whitham (1974) suggests introducing nondimensionalised quantities

$$q = \eta p, \quad Q(q) = \eta V P(p), \quad m = \frac{x}{Vt}. \tag{2.92}$$

In general, the boundary condition $h(t)$ will introduce another time scale T, say, and $A(p)$ will take the form

$$A(p) = \mathcal{A}\left(q\frac{T}{\eta}\right). \tag{2.93}$$

Then (2.88) becomes

$$c(x,t) = \frac{1}{2\pi i}\int_{l-i\infty}^{l+i\infty} \frac{\mathcal{A}(qT/\eta)}{q} \exp\left\{(q + mQ(q))\frac{t}{\eta}\right\} \mathrm{d}q, \tag{2.94}$$

where

$$Q(q) = \frac{Vq(1+q)}{Vq + V'}. \tag{2.95}$$

We now consider the asymptotic behaviour of (2.94) as $t/\eta \to \infty$ with m fixed. The dominant contribution comes from the neighbourhood of the point $q = q^\star$ for which

$$\frac{\mathrm{d}}{\mathrm{d}q}(q + mQ) = 0,$$

that is

$$1 + mQ'(q^\star) = 0. \tag{2.96}$$

This comes from the saddle point method (Ablowitz & Fokas 1997). The first term of the asymptotic expansion is found by deforming the contour into the path of steepest descent $\mathcal{C}$ through $q = q^\star$ and expanding $q + mQ$ as far as the quadratic terms. So, define

$$\Gamma(q) = q + mQ(q). \tag{2.97}$$

Then, a Taylor series expansion gives

$$\Gamma(q) = \Gamma(q^\star) + (q - q^\star)\Gamma'(q^\star) + \tfrac{1}{2}(q - q^\star)^2\Gamma''(q^\star) + \ldots.$$

Using (2.96) we have

$$\begin{aligned} \Gamma'(q^\star) &= 1 + mQ'(q^\star) = 0 \\ \Gamma''(q^\star) &= mQ''(q^\star), \end{aligned}$$

and so

$$\exp\left\{\left(q+mQ(q)\right)\frac{t}{\eta}\right\} \sim \exp\left\{\left(q^\star+mQ(q^\star)\right)\frac{t}{\eta}\right\}\exp\left\{\frac{1}{2}(q-q^\star)^2 mQ''(q^\star)\frac{t}{\eta}\right\}+\ldots.$$

Hence, (2.94) becomes

$$c(x,t) \sim \exp\left\{\left(q^\star+mQ(q^\star)\right)\frac{t}{\eta}\right\}\frac{1}{2\pi i}\int_C \frac{\mathcal{A}(qT/\eta)}{q}\exp\left\{\frac{1}{2}(q-q^\star)^2 mQ''(q^\star)\frac{t}{\eta}\right\}\,\mathrm{d}q, \tag{2.98}$$

as $t/\eta \to \infty$. The remaining part of the integral would then also be expanded in a Taylor series about $q = q^\star$ and $\mathcal{A}(qT/\eta)/q$ would be replaced by $\mathcal{A}(q^\star T/\eta)/q^\star$. This is valid for the limit $t/\eta \to \infty$, with T/η fixed and is relevant when $t >> \eta$ and $t >> T$.

However, we are interested in $t >> \eta$ and $t >> T$, independent of the size of t/T. So, we need to be more general. Let us convert (2.98) into the original variables. Then

$$c(x,t) \sim \exp\left\{tp^\star + xP(p^\star)\right\}\frac{1}{2\pi i}\int_C \frac{A(p)}{p}\exp\left\{\tfrac{1}{2}x(p-p^\star)^2 mP''(p^\star)\right\}\,\mathrm{d}p, \tag{2.99}$$

where $p^\star$ is the function of x and t which is determined by

$$t + xP'(p^\star) = 0. \tag{2.100}$$

This provides the asymptotic behaviour of c as $t/\eta \to \infty$ keeping $x/(Vt)$ fixed. For simplicity, we assume $\int_0^\infty h(t)\,\mathrm{d}t$ is convergent and so $A(p)/p$ is finite as $p \to 0$ and there is no pole. The expression in (2.99) is then dominated by the exponential outside the integral. The exponent is stationary when

$$\frac{\partial}{\partial x}\left\{tp^\star + xP(p^\star)\right\} = 0,$$

which reduces to (since $p^\star$ is a function of x and using (2.100))

$$P(p^\star) = 0.$$

From (2.86) this occurs when either $p^\star = 0$ or $p^\star = -1/\eta$. However, suppose we consider the exponential term outside the integral in (2.99). If $p^\star = 0$ then

$$\exp\left\{tp^\star + xP(p^\star)\right\} = 1,$$

whereas, if $p^\star = -1/\eta$, then

$$\exp\left\{tp^\star + xP(p^\star)\right\} = \exp\left\{-t/\eta\right\},$$

which will very quickly damp the solution as t/η becomes large. Hence $p^\star = 0$ is the correct choice for P and the maximum is found on

$$t + P'(0)x = 0.$$

A simple calculation using (2.86) gives $P'(0) = -1/V'$ and so the maximum of the exponential factor is found on

$$x = V't. \tag{2.101}$$

The disturbance is exponentially small (in this limit) except in the neighbourhood of $x = V't$. This result shows that the main part of the disturbance eventually travels with velocity V'. Since the approximation is for $t >> \eta$, the result applies increasingly earlier as $\eta \to 0$.

Further information can be deduced about the behaviour of the main disturbance. Suppose we consider (2.88) and expand $P(p)$ about $p = 0$. Now

$$P(p) = P(0) + pP'(0) + \tfrac{1}{2}p^2P''(0) + \ldots,$$

and, since $P(0) = 0$, $P'(0) = -1/V'$ and $P''(0) = 2\eta(V - V')/V'^2$ this leads to

$$P(p) \sim -\frac{p}{V'} + \frac{p^2\eta(V - V')}{V'^2} + \ldots.$$

Hence, in the neighbourhood of $x - V't = 0$ as $t/\eta \to \infty$ we can deduce that

$$c \sim \frac{1}{2\pi i}\int_C \frac{A(p)}{p}\exp\left\{p\left(t - \frac{x}{V'}\right) + \frac{p^2\eta(V - V')x}{V'^2}\right\}\,\mathrm{d}p. \tag{2.102}$$

The first approximation is simply

$$c \sim \frac{1}{2\pi i}\int_C \frac{A(p)}{p}\exp\left\{p\left(t - \frac{x}{V'}\right)\right\}\,\mathrm{d}p = h\left(t - \frac{x}{V'}\right),$$

which is exactly the prediction of the lower order equation (2.79). To see the effect of the quadratic term in the exponential in (2.102), it is more useful to find the equation satisfied by (2.102) rather than to interpret the integral. Suppose we have the second order equation

$$c_t + V'c_x = \frac{\eta(V - V')}{V'}c_{tt}, \tag{2.103}$$

with the same boundary condition $c = h(t)$ on $x = 0$. Following the same procedure as for (2.77) we apply the Heaviside transformation. Then

$$H(x,p) + V'\frac{1}{p}H_x(x,p) = \frac{\eta(V - V')}{V'}pH(x,p),$$

and the solution is

$$H(x,p) = A(p)e^{xR(p)}, \tag{2.104}$$

where

$$R(p) = -\frac{p}{V'}\left[1 - \frac{\eta(V-V')}{V'}p\right]. \tag{2.105}$$

Hence the solution of (2.103) is

$$\begin{aligned} c(x,t) &= \frac{1}{2\pi i}\int_{l-i\infty}^{l+i\infty} \frac{A(p)}{p}\exp\{pt + R(p)x\}\,\mathrm{d}p \\ &= \frac{1}{2\pi i}\int_{l-i\infty}^{l+i\infty} \frac{A(p)}{p}\exp\left\{p\left(t - \frac{x}{V'}\right) + \frac{p^2\eta(V-V')x}{V'^2}\right\}\mathrm{d}p, \end{aligned} \tag{2.106}$$

which is precisely (2.102). Let us examine (2.103) more closely. The right hand side is already a small correction (of order η/t compared with the other terms), so it is consistent to use the first approximation $\partial/\partial t \simeq -V'(\partial/\partial x)$ on c_{tt} to obtain

$$c_t + V'c_x = \eta(V-V')V'c_{xx}. \tag{2.107}$$

This shows that the main part of the disturbance propagates with velocity V' and is diffused by the effects of the higher order terms in the equation. The latter effect is small when η is small. Using the definitions of η and V' from (2.83), this becomes

$$c_t + \frac{V\mu}{\lambda+\mu}c_x = \frac{V^2\lambda\mu}{(\lambda+\mu)^3}c_{xx}, \tag{2.108}$$

which is precisely the Improved-equilibrium model derived in the previous Section.

Lastly, Lighthill & Whitham (1955) deduce a formula which gives the position of maximum depth of the main part of the disturbance for large t. This uses a standard formula from the saddle-point method to obtain an approximation to (2.88) when t is large. Hence we have

$$c \sim \frac{1}{\sqrt{(2\pi|P''(p^\star)|x)}}\frac{A(p^\star)}{p^\star}e^{p^\star t + P(p^\star)x}, \tag{2.109}$$

where $p^\star$ is the solution of (2.100). Since $p^\star = 0$ we substitute this into (2.109). As mentioned before we assume that $\int_0^\infty h(t)\,\mathrm{d}t$ is convergent and replace $A(p^\star)/p^\star$ with

$$\lim_{p\to 0}\frac{A(p)}{p} = \int_0^\infty h(t)\,\mathrm{d}t.$$

Then, since the exponential term is 1, (2.109) reduces to

$$c_{\max} \sim \sqrt{\frac{V'^2}{4\pi\eta(V-V')x}} \int_0^\infty h(t)\,\mathrm{d}t. \tag{2.110}$$

c attains its maximum when $x = V't$ (see (2.101)) and so we can substitute this into (2.110) and use the definition of V' to obtain

$$c_{\max} \sim \sqrt{\frac{\mu(\lambda+\mu)}{4\pi\lambda t}} \int_0^\infty h(t)\,\mathrm{d}t. \tag{2.111}$$

Suppose $g(t)$, the boundary condition for a, is a square pulse of area 1, i.e.

$$g(t) = \begin{cases} 1 & \text{if } 0 \le t \le 1 \\ 0 & \text{if } t > 1. \end{cases} \tag{2.112}$$

A simple calculation gives

$$\int_0^\infty h(t)\,\mathrm{d}t = 1 + \frac{\mu}{\lambda}, \tag{2.113}$$

and so, the expression in (2.111) becomes

$$c_{\max} \sim \xi := \frac{1}{2}\sqrt{\frac{(\lambda+\mu)^3}{\pi\mu\lambda t}}. \tag{2.114}$$

Table 2.1 shows comparisons of (2.114) with the exact maximum value of c (found from the Laplace transform solution when x is fixed). The expression in (2.114) is denoted by ξ and the final column shows the difference between these two values of $c_{\max}$ (i.e. the error). The value of t used to find ξ is taken to be x/V' as we have shown that c attains its maximum when $x = V't$ (see (2.101)). Note that V is always assumed to be 1. As t increases we can see that this reduces for all values of λ and μ. This confirms that our estimate is accurate for large values of t.

We finally summarise the features we have shown above. The first signals propagate out with velocity V but are damped, as shown in (2.91). The main disturbance lags behind and moves with the lower order wave speed V'. After a time of order η, the first signals are exponentially small and the main part of the solution to (2.77) is well described by (2.79) using the same boundary condition at $x = 0$. The effect of the higher order terms is to produce a diffusion of the lower order waves as shown by (2.108) but this is small when η is appropriately small.

$\lambda = 1$ and $\mu = 1$ (and $V' = 1/2$)				$\lambda = 9$ and $\mu = 1$ (and $V' = 1/10$)			
x	$\max(c)$	ξ	$\lvert\max(c) - \xi\rvert$	x	$\max(c)$	ξ	$\lvert\max(c) - \xi\rvert$
1	1.0009	0.5642	0.4637	1	0.9525	0.9403	0.0122
2	0.5820	0.3989	0.1831	2	0.6691	0.6649	0.0042
3	0.3496	0.3257	0.0239	3	0.5452	0.5429	0.0023
4	0.2965	0.2821	0.0144	4	0.4716	0.4702	0.0014
5	0.2626	0.2523	0.0103	5	0.4216	0.4205	0.0011
$\lambda = 45$ and $\mu = 5$ (and $V' = 1/5$)				$\lambda = 90$ and $\mu = 10$ (and $V' = 1/10$)			
x	$\max(c)$	ξ	$\lvert\max(c) - \xi\rvert$	x	$\max(c)$	ξ	$\lvert\max(c) - \xi\rvert$
1	2.0851	2.1026	0.0175	1	2.9836	2.9735	0.0101
2	1.4806	1.4868	0.0062	2	2.1096	2.1026	0.0070
3	1.2106	1.2139	0.0033	3	1.7190	1.7168	0.0023
4	1.0492	1.0513	0.0021	4	1.4890	1.4868	0.0022
5	0.9388	0.9403	0.0015	5	1.3316	1.3298	0.0017

Table 2.1: *Table showing the comparison of the estimate for c_{max} from (2.114) with the exact value for various λ and μ.*

2.7 Exponential decay of simple nonlinear models

We have already seen in this Chapter (and discussed in the Introduction) that the retarded speed is a very important feature of coupled reactive transport models. For the Linear Model we have been able to show the role of the reduced speed V' using various analyses: by solving the equations exactly using Laplace Transforms, by performing an asymptotic analysis and by using techniques similar to those in (Lighthill & Whitham 1955). We know that this phenomenon must be a feature of nonlinear models.

The general two system reactive transport model that we consider is (as defined in the Introduction, but which we state here again for convenience)

$$a_t + V a_x = f(a, b) \tag{2.115}$$

$$b_t = -f(a, b). \tag{2.116}$$

Our objective is to prove that bounds exist on the speed of propagation for the system (2.115) and (2.116) provided certain bounds exist on the partial derivatives of f. We need to show that the solution decays exponentially as a consequence of these assumptions before deducing the bounds on the retardation speed. This result is proved here and then used for the analysis in Section 2.8. Firstly, we restrict attention to the linearised problem; once this has been proved we can then deduce exponential decay of the solution for the nonlinear case using Picard iteration.

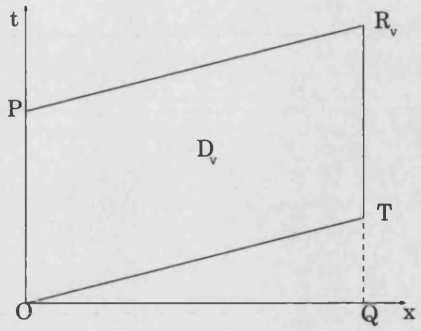

Figure 2-4: *The domain D_V in the (x,t)-plane.*

2.7.1 The linearised problem

In Section 2.6.2 we expressed the Linear Model as a single second order equation for c (see (2.76)). The same equation holds for the concentration a. We now suppose that λ and μ are functions of x and t and consider the domain D_V as shown in Figure 2-4. Note that we only need to integrate over OTR_VP, and not OQR_VP, since the solution is zero below the line OT provided we have zero initial data. This can easily be seen from examining the characteristics of the conservation law $c_t + Va_x = 0$: below the line $t = \frac{1}{V}x$ the solution emanates from the $t = 0$ axis and so is zero. The result is stated in the following Theorem:

Theorem 1. *Consider the linearised problem on the domain $D_V = OTR_VP$ (as shown in Figure 2-4) written in the form*

$$a_{tt} + Va_{xt} = -(\lambda + \mu)a_t - \mu V a_x, \tag{2.117}$$

with $\lambda = \lambda(x,t)$, $\mu = \mu(x,t)$ such that

$$\lambda \geq \lambda_0 > 0, \qquad \mu \geq \mu_0 > 0, \tag{2.118}$$

and which are uniformly bounded. Define initial and boundary data by

$$a(x,0) = a_x(x,0) = a_t(x,0) = 0, \tag{2.119}$$

Let P, T and R_V have co-ordinates $(0,t_P)$, $(x_Q, \frac{1}{V}x_Q)$ and $(x_Q, \frac{1}{V}x_Q + t_P)$ respectively. (2.120)

Then

$$a|_{R_V} \longrightarrow 0, \quad as \quad x_Q \longrightarrow \infty. \tag{2.121}$$

tively.

Then

$$a|_{R_V} \longrightarrow 0, \quad as \quad x_Q \longrightarrow \infty. \tag{2.121}$$

Proof. To solve (2.117) on D_V we first change variables to convert the problem into

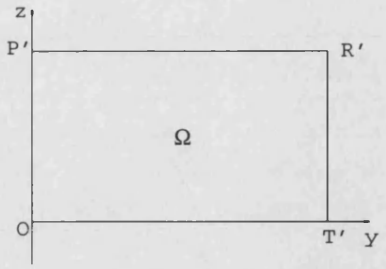

Figure 2-5: *The domain Ω in the (y,z)-plane.*

normal form. Define

$$y = \tfrac{1}{V}x, \qquad z = t - \tfrac{1}{V}x. \tag{2.122}$$

Then (2.117) reduces to the problem

$$L[a] \equiv a_{yz} + \mu a_y + \lambda a_z = 0, \tag{2.123}$$

where the domain D_V becomes the closed domain $\Omega = OT'R'P'$ (see Figure 2-5). The points P', T' and R' have co-ordinates $(0, t_P)$, $(\frac{1}{V}x_Q, 0)$ and $(\frac{1}{V}x_Q, t_P)$ respectively. In this proof we will use a Riemann's function (see (Guenther & Lee 1988, pages 129-132) and (Garabedian 1964, pages 127-134)) to find the solution in terms of definite integrals. The adjoint operator $M[v]$ is defined as

$$M[v] \equiv v_{yz} - (\mu v)_y - (\lambda v)_z, \tag{2.124}$$

and so

$$vL[a] - aM[v] \;=\; (\mu va - v_z a)_y + (\lambda va + va_y)_z. \tag{2.125}$$

We can now integrate both sides of the above expression over Ω and apply the Divergence Theorem (Spiegel 1959, page 106) which, in general, is given by

$$\iint_D (N_x - M_t)\,\mathrm{d}x\,\mathrm{d}t = \oint_{\partial D} \left[M\,\mathrm{d}x + N\,\mathrm{d}t\right], \tag{2.126}$$

where D is traversed in the positive (anti-clockwise) direction. Hence (2.125) becomes

$$\begin{aligned} \iint_\Omega (vL[a] - aM[v])\,\mathrm{d}y\,\mathrm{d}z &= \oint_{\partial\Omega} \left[\mu va - v_z a\right]\mathrm{d}z - \left[\lambda va + va_y\right]\mathrm{d}y \\ &= \oint_{\partial\Omega} \left[\mu v - v_z\right] a\,\mathrm{d}z - \oint_{\partial\Omega} \left[\lambda va - av_y + (va)_y\right]\mathrm{d}y. \end{aligned}$$

Since $a = 0$ along OT' (from the first condition in (2.119)), the above reduces to

$$\iint_{\Omega} (vL[a] - aM[v])\,\mathrm{d}y\,\mathrm{d}z = \int_{T'}^{R'} [\mu v - v_z]a\,\mathrm{d}z - \int_{O}^{P'} [\mu v - v_z]a\,\mathrm{d}z + \int_{P'}^{R'} [\lambda v - v_y]a\,\mathrm{d}y + (av)|_{R'} - (av)|_{P'}. \quad (2.127)$$

We now choose v to satisfy the following (and so eliminating two of the integrals above):

$$M[v] \equiv v_{yz} - (\mu v)_y - (\lambda v)_z = 0 \quad (2.128)$$

$$v_z = \mu v \quad \text{on} \quad T'R' \quad (2.129)$$

$$v_y = \lambda v \quad \text{on} \quad P'R' \quad (2.130)$$

$$v = 1 \quad \text{at} \quad R'. \quad (2.131)$$

Since $L[a] = 0$ (see (2.123)), we can substitute this, along with the equations (2.128)–(2.131), into (2.127) to obtain

$$a|_{R'} = (av)|_{P'} + \int_{O}^{P'} [\mu v - v_z]a\,\mathrm{d}z. \quad (2.132)$$

Now

$$\int_{O}^{P'} [\mu v - v_z]a\,\mathrm{d}z = -(av)|_{P'} + \int_{O}^{P'} [\mu a + a_z]v\,\mathrm{d}z.$$

Hence, (2.132) simplifies to

$$a|_{R'} = \int_{O}^{P'} [\mu a + a_z]v\,\mathrm{d}z. \quad (2.133)$$

Using the boundary data (2.120) this can be written as

$$a|_{R'} = \int_{O}^{P'} \left[\mu g\left(\frac{z}{V}\right) + g'\left(\frac{z}{V}\right)\right] v\,\mathrm{d}z. \quad (2.134)$$

Suppose the point R' has co-ordinates (ξ, η). The problem to solve for v is given by (2.128)–(2.131) and is known as the Goursat problem (Garabedian 1964, pages 117-119). The conditions (2.129) and (2.130) are actually ordinary differential equations for v along $P'R'$ and $T'R'$ respectively. These can be solved to give

$$v(y, \eta) = \exp\left[\int_{\xi}^{y} \lambda(\sigma, \eta)\,\mathrm{d}\sigma\right] \quad (2.135)$$

$$v(\xi, z) = \exp\left[\int_{\eta}^{z} \mu(\xi, \sigma)\,\mathrm{d}\sigma\right]. \quad (2.136)$$

Assuming that λ and μ are strictly positive for all (y, z) in the domain Ω, the integrals in these exponentials will be negative (since we are also assuming $y < \xi$ and $z < \eta$). Hence, from the point R', the variable v decays back exponentially along $P'R'$ and $T'R'$ towards the axes. We wish to show that v is exponentially small along OP'; then we can deduce that a tends to zero as ξ tends to infinity using (2.134).

The solution of the Goursat problem (2.128)–(2.131) over the domain Ω is given by

$$v|_{R'} = v|_{P'} - v|_O + v|_{T'} + \iint_\Omega \left[(\mu v)_y + (\lambda v)_z\right] \mathrm{d}y\,\mathrm{d}z. \tag{2.137}$$

We do not restrict the left hand corner of Ω to be at the origin, but instead label this S'. Then T' is the corresponding point horizontal to S' (and so is not necessarily on the y axis), and (2.137) can be rearranged to give

$$v|_S = v|_P - v|_R + v|_T + \iint_\Omega \left[(\mu v)_y + (\lambda v)_z\right] \mathrm{d}y\,\mathrm{d}z, \tag{2.138}$$

where, for convenience, we have dropped the $(\cdot)'$ notation. The term under the double integral sign is in divergence form so we can apply (2.126). This leads to

$$v|_S = v|_R - \int_S^P \mu v\,\mathrm{d}z - \int_S^T \lambda v\,\mathrm{d}y, \tag{2.139}$$

using the fact that $v_z = \mu v$ along TR and $v_y = \lambda v$ along PR. Suppose S has co-ordinates (y, z). Then (2.139) becomes

$$v(y,z) = 1 - \int_z^\eta \mu(y,z')v(y,z')\,\mathrm{d}z' - \int_y^\xi \lambda(y',z)v(y',z)\,\mathrm{d}y'. \tag{2.140}$$

We would like to deduce that v is bounded by a negative exponential. Unfortunately, this cannot easily be done using the expression (2.140) directly. Firstly, we prove a Lemma which is the opposite to the situation for v: the solution is known on the left and bottom sides of the rectangle and exponential decay is proved out towards infinity. It can then easily be adapted to the Goursat problem. This is analogous to the following result for a simple ODE:

Lemma 1. *Consider $y'(x) = \lambda(x)y(x)$, with $y(\xi) = 1$. If $\lambda(x) \leq -\lambda_0 < 0$ and $|\lambda|$ is uniformly bounded then y decays at least as fast as $e^{-\lambda_0(y-\xi)}$.*

This is proved in Appendix A (Section A.4). Now consider the following Lemma:

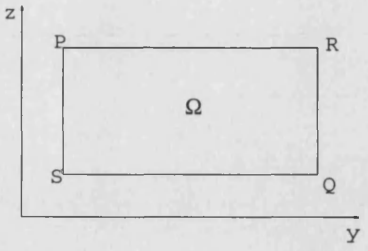

Figure 2-6: *The domain Ω in the (y,z)-plane for the Goursat problem.*

Lemma 2. *Consider $v:\ \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ and the domain Ω (see Figure 2-6) such that*

$$v_{yz} = (\mu v)_y + (\lambda v)_z, \qquad (y,z) \in \Omega \tag{2.141}$$

$$v_z = \mu v \quad \textit{along } SP \tag{2.142}$$

$$v_y = \lambda v \quad \textit{along } SQ \tag{2.143}$$

$$v = 1 \quad \textit{at } S, \tag{2.144}$$

where

$$\lambda \leq -\lambda_0 < 0, \quad \mu \leq -\mu_0 < 0, \tag{2.145}$$

and λ and μ are uniformly bounded for all $(y,z) \in \Omega$. Suppose S has co-ordinates (ζ,ρ). Then

$$\left|v(y,z)e^{\lambda_0(y-\zeta)+\mu_0(z-\rho)}\right| \leq K, \quad \forall\ (y,z) \in \Omega, \tag{2.146}$$

for some constant $K > 0$.

Proof. To prove Lemma 2 we define

$$w(y,z) = v(y,z)e^{(\lambda_0+\gamma)(y-\zeta)+(\mu_0+\delta)(z-\rho)}, \tag{2.147}$$

where γ and δ are positive constants, and show that (2.146) holds in an appropriately defined norm. A simple calculation gives

$$w_{yz} = \left[\mu w+(\mu_0+\delta)w\right]_y + \left[\lambda w+(\lambda_0+\gamma)w\right]_z + \left[\mu\lambda-(\mu+\mu_0+\delta)(\lambda+\lambda_0+\gamma)\right]w. \tag{2.148}$$

Along SP and SQ we know that (2.142) and (2.143) hold respectively. In terms of w these ODEs become

$$w_z = (\mu+\mu_0+\delta)w \quad \text{along } SP \tag{2.149}$$

$$w_y = (\lambda+\lambda_0+\gamma)w \quad \text{along } SQ. \tag{2.150}$$

Also, $w = 1$ at S. Integrating (2.148) over Ω, and then applying the Divergence theorem

to the first two terms, gives

$$\begin{aligned} w|_R &= w|_S + \int_Q^R (\mu+\mu_0+\delta)w\,\mathrm{d}z + \int_P^R (\lambda+\lambda_0+\gamma)w\,\mathrm{d}y \\ &\quad + \iint_\Omega \left[\mu\lambda - (\mu+\mu_0+\delta)(\lambda+\lambda_0+\gamma)\right]w\,\mathrm{d}y\,\mathrm{d}z. \end{aligned} \tag{2.151}$$

This can be written in the form of an operator equation by defining

$$\begin{aligned} (Tw)(y,z) &:= 1 + \int_\rho^z \left[\mu(y,z')+\mu_0+\delta\right]w(y,z')\,\mathrm{d}z' + \int_\zeta^y \left[\lambda(y',z)+\lambda_0+\gamma\right]w(y',z)\,\mathrm{d}y' \\ &\quad + \int_\rho^z \int_\zeta^y \Big[\mu(y',z')\lambda(y',z') - \big(\mu(y',z')+\mu_0+\delta\big)\big(\lambda(y',z')+\lambda_0+\gamma\big)\Big]w(y',z')\,\mathrm{d}y'\,\mathrm{d}z'. \end{aligned} \tag{2.152}$$

The integral operator T maps each function w from the Banach space $C(\Omega)$ of continuous functions to a function Tw in the same space. Since $|\lambda|$ is uniformly bounded we can assume the following:

$$0 \le \lambda+\lambda_0+\gamma \le \frac{1}{n}\gamma, \qquad 0 \le \mu+\mu_0+\delta \le \frac{1}{m}\delta, \tag{2.153}$$

for some constants n, $m > 1$. Also,

$$|\lambda| \le L, \quad |\mu| \le M, \qquad L,\ M > 0, \tag{2.154}$$

$\forall\ \lambda,\ \mu \in \Omega$. So

$$\left|\mu\lambda - (\mu+\mu_0+\delta)(\lambda+\lambda_0+\gamma)\right| \ \le\ LM + \frac{1}{n}\gamma\frac{1}{m}\delta. \tag{2.155}$$

Suppose u is another solution satisfying (2.152). Then

$$\begin{aligned} \left|(Tw)(y,z) - (Tu)(y,z)\right| \ \le\ & \frac{1}{m}\delta\int_\rho^z \left|w(y,z') - u(y,z')\right|\mathrm{d}z' \\ & + \frac{1}{n}\gamma\int_\zeta^y \left|w(y',z) - u(y',z)\right|\mathrm{d}y' \\ & + \left[LM + \frac{1}{n}\gamma\frac{1}{m}\delta\right]\int_\rho^z\int_\zeta^y \left|w(y',z') - u(y',z')\right|\mathrm{d}y'\,\mathrm{d}z'. \end{aligned} \tag{2.156}$$

Now define the norm

$$||w||_e = \max\left\{|w(y,z)|e^{-\alpha(y-\zeta)-\beta(z-\rho)} :\ (y,z) \in \Omega\right\}, \tag{2.157}$$

where α and β are positive constants. Then, inserting the exponential expression

$$e^{-\alpha(y-\zeta)-\beta(z-\rho)}e^{\alpha(y-\zeta)+\beta(z-\rho)},$$

inside each of the integrals in (2.156) gives

$$\begin{aligned}\left|(Tw)(y,z)-(Tu)(y,z)\right| \;\leq\; & \frac{1}{m}\delta||w-u||_e\int_\rho^z e^{\alpha(y-\zeta)+\beta(z'-\rho)}\,\mathrm{d}z' \\ & +\frac{1}{n}\gamma||w-u||_e\int_\zeta^y e^{\alpha(y'-\zeta)+\beta(z-\rho)}\,\mathrm{d}y' \\ & +\left[LM+\frac{1}{n}\gamma\frac{1}{m}\delta\right]||w-u||_e\int_\rho^z\int_\zeta^y e^{\alpha(y'-\zeta)+\beta(z'-\rho)}\,\mathrm{d}y'\,\mathrm{d}z'.\end{aligned}$$

This leads to

$$||Tw-Tu||_e\leq\left[\frac{1}{m\beta}\delta+\frac{1}{n\alpha}\gamma+\frac{LM}{\alpha\beta}+\frac{1}{m\beta}\delta\frac{1}{n\alpha}\gamma\right]||w-u||_e. \tag{2.158}$$

Hence, provided

$$\frac{1}{m}\delta<\frac{1}{N_\mu}\beta,\qquad\frac{1}{n}\gamma<\frac{1}{N_\lambda}\alpha, \tag{2.159}$$

for some N_μ, $N_\lambda>1$, and L and M are chosen appropriately, we can ensure

$$\frac{1}{m\beta}\delta+\frac{1}{n\alpha}\gamma+\frac{LM}{\alpha\beta}+\frac{1}{m\beta}\delta\frac{1}{n\alpha}\gamma<1. \tag{2.160}$$

Thus T satisfies a Lipschitz condition and is a contraction (see (Walter 1998, page 59) for details). Then w is bounded in the norm (2.157), and so

$$|w(y,z)|e^{-\alpha(y-\zeta)-\beta(z-\rho)}\leq K, \tag{2.161}$$

where $K>0$ is constant. Substituting v from (2.147) into this expression leads to

$$|v(y,z)|e^{(\lambda_0+\gamma-\alpha)(y-\zeta)+(\mu_0+\delta-\beta)(z-\rho)}\leq K. \tag{2.162}$$

We require $\gamma-\alpha\geq 0$ and $\delta-\beta\geq 0$ to deduce

$$|v(y,z)|e^{\lambda_0(y-\zeta)+\mu_0(z-\rho)}\leq K. \tag{2.163}$$

Combining these inequalities with (2.159) gives

$$\frac{N_\lambda}{n}\gamma<\alpha\leq\gamma,\qquad\frac{N_\mu}{m}\delta<\beta\leq\delta. \tag{2.164}$$

Finally, we must restrict $N_\lambda<n$ and $N_\mu<m$. The proof of Lemma 2 is complete. □

We now apply Lemma 2 to the Goursat problem (2.128)–(2.131) to complete the proof of Theorem 1. Firstly, the problem must be converted so the known value of v is at the bottom left corner of the domain. Hence set

$$\bar{y} = \xi - y, \qquad \bar{z} = \eta - z. \tag{2.165}$$

Then the problem becomes

$$v_{\bar{y}\bar{z}}(\bar{y},\bar{z}) = -(\mu v)_{\bar{y}}(\bar{y},\bar{z}) - (\lambda v)_{\bar{z}}(\bar{y},\bar{z}), \tag{2.166}$$

$$v_{\bar{y}}(\bar{y},0) = -\lambda v(\bar{y},0), \tag{2.167}$$

$$v_{\bar{z}}(0,\bar{z}) = -\mu v(0,\bar{z}), \tag{2.168}$$

$$v(0,0) = 1. \tag{2.169}$$

This is now in the form of Lemma 2 where λ and μ are replaced by $-\lambda$ and $-\mu$ and ζ and ρ are set to zero. Hence, provided

$$-\lambda \leq -\lambda_0 < 0, \qquad -\mu \leq -\mu_0 < 0, \tag{2.170}$$

and $|\lambda|$ and $|\mu|$ are uniformly bounded then

$$|v(\bar{y},\bar{z})|e^{\lambda_0\bar{y}+\mu_0\bar{z}} \leq K. \tag{2.171}$$

Finally, using (2.165), we can deduce that v is exponentially decreasing, i.e.

$$|v(y,z)|e^{\lambda_0(\xi-y)+\mu_0(\eta-z)} \leq K. \tag{2.172}$$

In particular

$$|v(0,z)| \leq Ke^{-\lambda_0\xi-\mu_0(\eta-z)}. \tag{2.173}$$

We have shown that v is exponentially small along OP' and so, using (2.133), the result in Theorem 1 can be deduced. Hence the proof is complete. □

2.7.2 The nonlinear problem

We can finally prove the same result for a general nonlinear reaction term f provided certain bounds hold on the partial derivatives.

Theorem 2. *The nonlinear model (2.115) and (2.116) can be written as the following second order equation:*

$$a_{tt} + Va_{xt} = (f_a - f_b)a_t - Vf_b a_x, \tag{2.174}$$

where $f_a = \partial f/\partial a$, $f_b = \partial f/\partial b$ *and*

$$f_a \leq -\lambda_0 < 0, \qquad -f_b \leq -\mu_0 < 0. \tag{2.175}$$

Then

$$a|_{R_V} \longrightarrow 0, \quad as \quad x_Q \longrightarrow \infty. \tag{2.176}$$

Proof. Consider the transport equation (2.115) and differentiate this with respect to t. Then

$$a_{tt} + V a_{xt} = f_a a_t + f_b b_t. \tag{2.177}$$

Now, adding (2.115) and (2.116) gives

$$b_t = -a_t - V a_x,$$

and so we can substitute this into (2.177) to eliminate the b_t term

$$a_{tt} + V a_{xt} = f_a a_t + f_b\big(-a_t - V a_x\big),$$

which is precisely (2.174). As in Theorem 1 we can change variables by setting $y = \frac{1}{V}x$ and $z = Vt - x$. Then (2.174) becomes

$$a_{yz} = -f_b a_y + f_a a_z, \tag{2.178}$$

and the domain D_V is now the rectangle Ω (see Figure 2-5). The partial derivatives f_a and f_b may depend on a and b, and so the adjoint operator cannot be found directly. However, we can consider the iterated equation

$$a_{yz}^{(n+1)} = -f_b^{(n)} a_y^{(n+1)} + f_a^{(n)} a_z^{(n+1)}, \tag{2.179}$$

which is linear for each iteration n, and with $b^{(n+1)}$ being given by the (unique) solution of $b_t = -f\big(a^{(n+1)}, b\big)$. Equation (2.179) is now a linear problem for $a^{(n+1)}$; we can therefore apply Theorem 1. Hence $a^{(n+1)}$ decays exponentially provided $f_b^{(n)}$ and $f_a^{(n)}$ are uniformly bounded and satisfy (2.175). The proof of Theorem 1 indicates a contraction argument, and so, by the Contraction Mapping Principle (Walter 1998, page 59), one can show that $a^{(n+1)}$ converges to a in the norm (2.157). Finally, we deduce that the result (2.176) holds. □

2.8 Conservation properties

In this Section we are able to deduce bounds on the speed of propagation for a general nonlinear reaction term. To achieve this we consider a general domain $D_S := OTR_SP$ (as shown in Figure 2-7) which is identical to D_V in Figure 2-4 except that $\mathrm{d}x = S\mathrm{d}t$

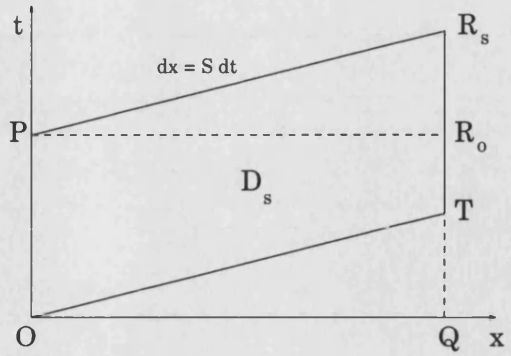

Figure 2-7: *The domain D_S in the (x,t)-plane.*

along the line PR_S. The conservation law $c_t + Va_x = 0$ is integrated over the domain D_S which gives an integral expression relating the integrals along the boundary of D_S. We can use this to obtain the required result.

2.8.1 Integration over a general domain OTR_SP

Consider the general two equation reactive transport model (2.115) and (2.116). If these are added together we can eliminate the source term to obtain a conservation equation which will hold for any $f(a,b)$, i.e.

$$c_t + Va_x = 0, \tag{2.180}$$

where $c = a + b$. Consider the following result.

Lemma 3. *If the conservation law (2.180) is integrated over the domain D_S then*

$$\int_P^{R_S} \left(c - \frac{V}{S}a\right) \mathrm{d}x = V\left[\int_O^P a\,\mathrm{d}t - \int_T^{R_S} a\,\mathrm{d}t\right]. \tag{2.181}$$

Proof. Integrating (2.180) over D_S and applying the Divergence theorem (2.126) leads to

$$\int_{\partial D_S} \left[-c\,\mathrm{d}x + Va\,\mathrm{d}t\right] = 0. \tag{2.182}$$

Since $\mathrm{d}x = 0$ on OP and TR_S, and $a = c = 0$ along OT, this becomes

$$\int_P^{R_S} c\,\mathrm{d}x + V\int_T^{R_S} a\,\mathrm{d}t - V\int_P^{R_S} a\,\mathrm{d}t - V\int_O^P a\,\mathrm{d}t = 0. \tag{2.183}$$

We also know that $\mathrm{d}t = \frac{1}{S}\mathrm{d}x$ along PR_S and so (2.183) can be rearranged to give (2.181). □

N.B. Suppose we consider the point R_O which has co-ordinates (x_Q, t_P). If $R_S \to R_O$ and $T \to Q$ then the domain D_S becomes $D_O := OQR_OP$ (see Figure 2-7). This means that $1/S \to 0$ and so, in the limit, (2.181) becomes

$$\int_P^{R_O} c\,\mathrm{d}x = V\left[\int_O^P a\,\mathrm{d}t - \int_T^{R_O} a\,\mathrm{d}t\right], \tag{2.184}$$

which is the expected result of integrating a conservation law over a box.

The integral expression in Lemma 3 will be the basis of our analysis to obtain bounds on S.

2.8.2 Extending the domain to infinity

Consider the integral of a along the line TR_S that appears in the expression (2.181) from Lemma 3. We first state some properties of this integral and then prove that it must decay to zero as the domain is extended to infinity (for a general reaction equation (2.116)). Define

$$A(x_Q, S) := \int_T^{R_S} a(x_Q, t)\,\mathrm{d}t. \tag{2.185}$$

Then

$$A(x_Q, S) = \int_0^{\frac{1}{S}x_Q + t_P} a(x_Q, t)\,\mathrm{d}t = \int_{\frac{1}{V}x_Q}^{\frac{1}{S}x_Q + t_P} a(x_Q, t)\,\mathrm{d}t, \tag{2.186}$$

which holds since a is zero below the line $t = \frac{1}{V}x$. Also

$$A(0, S) = \int_O^P a(0, t)\,\mathrm{d}t > 0. \tag{2.187}$$

Hence we can rearrange (2.181) to give an expression for $A(x_Q, S)$, namely

$$\begin{aligned} A(x_Q, S) &= A(0, S) - \int_P^{R_S} \left(\frac{c}{V} - \frac{a}{S}\right) \mathrm{d}x \\ &= A(0, S) - \int_0^{x_Q} \left(\frac{c}{V} - \frac{a}{S}\right)\Big|_{x = S(t - t_P)} \mathrm{d}x, \end{aligned} \tag{2.188}$$

since $x = S(t - t_P)$ along PR_S. Differentiating (2.188) with respect to x gives

$$A_x(x_Q, S) = -\left(\frac{c}{V} - \frac{a}{S}\right)\Big|_{R_S}, \tag{2.189}$$

or, eliminating c $(= a + b)$, this becomes

$$A_x(x_Q, S) = -\left(\frac{1}{V} - \frac{1}{S}\right) a|_{R_S} - \frac{1}{V} b|_{R_S}. \tag{2.190}$$

We now consider the reaction term $f(a,b)$ and suppose the bounds on the partial derivatives in (2.175) hold. This can be written as $f(\mathbf{y})$ where $\mathbf{y} = (a,b)$. Then, by the Fundamental Theorem of line integrals, we can express f as the integral

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{0}) &= \int_0^1 \nabla f(t\mathbf{y}) \cdot \mathbf{y}\, \mathrm{d}t. \\ &= \int_0^1 \left[f_a(t\mathbf{y})a + f_b(t\mathbf{y})b \right] \mathrm{d}t. \end{aligned} \tag{2.191}$$

Assuming $f(0,0) = 0$ and imposing the extra bound

$$f_b \leq \mu_1, \tag{2.192}$$

with $\mu_1 > 0$, gives

$$f(a,b) \leq -\lambda_0 a + \mu_1 b. \tag{2.193}$$

Substituting this into (2.116) (i.e. $b_t = -f(a,b)$) and solving the resulting ODE for b leads to

$$b(x,t) \geq \lambda_0 \int_0^t a(x,r) e^{-\mu_1(t-r)}\, \mathrm{d}r. \tag{2.194}$$

This result will be used to prove $A(x_Q, S) \to 0$ as $x_Q \to \infty$. First suppose $S = V$. Then

Lemma 4. *$A(x_Q, V) \longrightarrow 0$ as $x_Q \longrightarrow \infty$.*

Proof. Setting $S = V$ in (2.190) eliminates the first term and so

$$A_x(x_Q, V) = -\frac{1}{V} b|_{R_V}.$$

Since $x = x_Q$ and $t = \frac{1}{V}x_Q + t_P$ at the point R_V, we can substitute the inequality (2.194) into the above expression for A_x to give

$$\begin{aligned} A_x(x_Q, V) &\leq -\frac{\lambda_0}{V} \int_0^{\frac{1}{V}x_Q + t_P} a(x_Q, r) e^{-\mu_1(\frac{1}{V}x_Q + t_P - r)}\, \mathrm{d}r \\ &= -\frac{\lambda_0}{V} \int_{\frac{1}{V}x_Q}^{\frac{1}{V}x_Q + t_P} a(x_Q, r) e^{-\mu_1(\frac{1}{V}x_Q + t_P - r)}\, \mathrm{d}r. \end{aligned} \tag{2.195}$$

Now

$$e^{-\mu_1(\frac{1}{V}x_Q + t_P - r)} \geq e^{-\mu_1 t_P},$$

since $\frac{1}{V}x_Q \leq r \leq \frac{1}{V}x_Q + t_P$. Hence (2.195) simplifies to

$$\begin{aligned} A_x(x_Q, V) &\leq -\frac{\lambda_0}{V} e^{-\mu_1 t_P} \int_{\frac{1}{V}x_Q}^{\frac{1}{V}x_Q + t_P} a(x_Q, r)\, \mathrm{d}r \\ &= -\frac{\lambda_0}{V} e^{-\mu_1 t_P} A(x_Q, V), \end{aligned} \tag{2.196}$$

which holds since $\lambda_0 > 0$. We can solve (2.196) to obtain

$$A(x_Q, V) \leq A(0, V) \exp\left\{-\frac{\lambda_0}{V} x_Q e^{-\mu_1 t_P}\right\}, \tag{2.197}$$

and the right hand side tends to zero as $x_Q \to \infty$. This completes the proof. □

We can now show that

$$A(x_Q, S) \longrightarrow 0 \quad \text{as} \quad x_Q \longrightarrow \infty, \tag{2.198}$$

provided $S > V$. To do this we use a comparison argument, i.e.

Lemma 5. *Consider two speeds S_1 and S_2 with $S_2 > S_1$. Then, if*

$$A(x_Q, S_1) \longrightarrow 0 \quad \textit{as} \quad x_Q \longrightarrow \infty, \tag{2.199}$$

it follows that

$$A(x_Q, S_2) \longrightarrow 0 \quad \textit{as} \quad x_Q \longrightarrow \infty. \tag{2.200}$$

Proof. From (2.186) we have

$$A(x_Q, S_2) = \int_0^{\frac{1}{S_2}x_Q + t_P} a(x_Q, t)\, \mathrm{d}t \leq \int_0^{\frac{1}{S_1}x_Q + t_P} a(x_Q, t)\, \mathrm{d}t = A(x_Q, S_1),$$

since $a(\cdot, \cdot) \geq 0$ and $S_2 > S_1$. □

We have shown that $A(x_Q, S) \to 0$ as $x_Q \to \infty$, for all $S \geq V$. This result is now proved for the situation when $S < V$.

Theorem 3.

$$A(x_Q, S) \longrightarrow 0, \quad \textit{as} \quad x_Q \longrightarrow \infty. \tag{2.201}$$

Proof. Following the proof of Lemma 4 we have, on substituting (2.194) into (2.190) and using the fact that $t = \frac{1}{S}x_Q + t_P$ at R_S

$$A_x(x_Q, S) \leq \left(\frac{1}{S} - \frac{1}{V}\right) a|_{R_S} - \frac{\lambda_0}{V} \int_{\frac{1}{V}x_Q}^{\frac{1}{S}x_Q + t_P} a(x_Q, r) e^{-\mu_1(\frac{1}{S}x_Q + t_P - r)}\, \mathrm{d}r. \tag{2.202}$$

Also, since $\frac{1}{V}x_Q \leq r \leq \frac{1}{S}x_Q + t_P$

$$e^{-\mu_1(\frac{1}{S}x_Q + t_P - r)} \geq e^{-\mu_1[(\frac{1}{V}-\frac{1}{S})x_Q - t_P]}.$$

Hence (2.202) becomes

$$A_x(x_Q, S) \leq \left(\frac{1}{S} - \frac{1}{V}\right) a|_{R_S} - \frac{\lambda_0}{V} \exp\left\{-\mu\left[\left(\frac{1}{S} - \frac{1}{V}\right) x_Q + t_P\right]\right\} A(x_Q, S). \quad (2.203)$$

In Section 2.7 we proved that $a_{R_V} \to 0$ as $x_Q \to \infty$. Since $S < V$ this also holds for a_{R_S}. The result in (2.201) now follows directly; it is straightforward to show that if

$$A'(x) \leq -\gamma A(x) + \psi(x), \quad (2.204)$$

with $\gamma > 0$, $A(0) > 0$ and $\psi(x) \to 0$ as $x \to \infty$, then

$$A(x) \longrightarrow 0 \quad \text{as} \quad x \longrightarrow \infty. \quad (2.205)$$

Applying this result to (2.203) with

$$\psi(x) = a|_{R_S}, \quad \gamma = \frac{\lambda}{V} \exp\left\{-\mu\left[\left(\frac{1}{S} - \frac{1}{V}\right) x_Q + t_P\right]\right\} > 0,$$

completes the proof. □

2.8.3 Lower bound for the reduced speed

We can now make deductions about the reduced speed using Theorem 3. Since $\mathrm{d}x = S\mathrm{d}t$ and $c = a + b$, the result (2.181) from Lemma 3 can be rewritten as

$$\frac{S}{V}\int_P^{R_S} b\,\mathrm{d}t + \left(\frac{S}{V} - 1\right)\int_P^{R_S} a\,\mathrm{d}t = \int_O^P a\,\mathrm{d}t - A(x_Q, S). \quad (2.206)$$

Assume that the boundary condition for a is an injection of a short pulse of chemical pollutants into the groundwater at $x = 0$, as defined in (2.24) in Section 2.2. If $\alpha = 1$ in this definition, then $\int_0^P a\,\mathrm{d}t = 1$ (provided P lies outside the region $[0, \delta]$). Suppose $x_Q \to \infty$ and define

$$\begin{aligned} \lim_{x_Q\to\infty}\int_P^{R_S} b\,\mathrm{d}t &= \lim_{x_Q\to\infty}\int_{t_P}^{t_P+\frac{1}{V}x_Q} b\big(S(t-t_P), t\big)\,\mathrm{d}t \\ &=: \int_{t_P}^{\infty} b\big(S(t-t_P), t\big)\,\mathrm{d}t. \end{aligned} \quad (2.207)$$

We can apply Theorem 3 to eliminate $A(x_Q, S)$. Hence (2.206) simplifies to

$$\frac{S}{V}\int_{t_P}^{\infty} b\big(S(t-t_P),t\big)\,\mathrm{d}t + \left(\frac{S}{V}-1\right)\int_{t_P}^{\infty} a\big(S(t-t_P),t\big)\,\mathrm{d}t = 1. \tag{2.208}$$

If $S = V$ then this relation becomes

$$\int_{t_P}^{\infty} b\big(V(t-t_P),t\big)\,\mathrm{d}t = 1. \tag{2.209}$$

Hence the solution must decay as $x_Q \to \infty$ and therefore no travelling wave solution can exist at this speed.

If $S > V$ then the left hand side of (2.208) is positive and equals 1; again no travelling wave solution can exist at this speed.

Lastly, suppose $S < V$. The second term on the left hand side of (2.208) is negative and so, potentially, a travelling wave solution could exist at this speed. However, provided we can bound the integral of b along this characteristic by the integral of a, we can show that a lower bound exists on the speed S. We require the following to hold:

Lemma 6.

$$\int_{t_P}^{\infty} b\big(S(t-t_P),t\big)\,\mathrm{d}t \le K\int_{t_P}^{\infty} a\big(S(t-t_P),t\big)\,\mathrm{d}t, \tag{2.210}$$

for some positive constant K.

Proof. To prove (2.210) we need to again consider f expressed as an integral involving its partial derivatives (i.e. (2.191)). Now impose the bound

$$-f_a \le \lambda_1, \tag{2.211}$$

with $\lambda_1 > 0$. Then

$$f(a,b) \ge -\lambda_1 a + \mu_0 b, \tag{2.212}$$

and so

$$b(x,t) \le \lambda_1 \int_0^t a(x,r)e^{-\mu_0(t-r)}\,\mathrm{d}r. \tag{2.213}$$

Hence

$$b\big(S(t-t_P),t\big) \le \lambda_1 \int_0^t a\big(S(t-t_P),r\big)e^{-\mu_0(t-r)}\mathrm{d}r. \tag{2.214}$$

We now state the *Second Mean Value Theorem of the Integral Calculus*, see (Courant 1934, pages 256-257), which can be used to simplify the integral in (2.214).

Theorem 4. *Suppose the function $\phi(t)$ is monotonic and continuous in the interval $t_1 \le t \le t_2$, and that the derivative $\phi'(t)$ is continuous. Further suppose that $f(t)$ is*

an arbitrary function continuous in the same interval. Then there exists a number τ, *such that* $t_1 \leq \tau \leq t_2$, *for which*

$$\int_{t_1}^{t_2} f(t)\phi(t)\,\mathrm{d}t = \phi(t_1)\int_{t_1}^{\tau} f(t)\,\mathrm{d}t + \phi(t_2)\int_{\tau}^{t_2} f(t)\,\mathrm{d}t. \tag{2.215}$$

We apply this to the integral in (2.214) over the interval $[0,t]$ where, in our case, ϕ is the function a and f is the exponential term $e^{-\mu_0(t-\tau)}$. Since a is zero when $t=0$, the first expression in (2.215) disappears and we have

$$\begin{aligned}\int_0^t a\big(S(t-t_P),r\big)e^{-\mu_0(t-r)}\mathrm{d}r &= a\big(S(t-t_P),t\big)\int_{\tau}^{t} e^{-\mu_0(t-r)}\mathrm{d}r \\ &= \frac{1}{\mu_0}a\big(S(t-t_P),t\big)\left[1-e^{-\mu_0(t-\tau)}\right] \\ &\leq \frac{1}{\mu_0}a\big(S(t-t_P),t\big),\end{aligned} \tag{2.216}$$

since $0 \leq \tau \leq t$. Substituting this into (2.214) leads to

$$b\big(S(t-t_P),t\big) \leq \frac{\lambda_1}{\mu_0}a\big(S(t-t_P),t\big). \tag{2.217}$$

Integrating this along PR_S as $x_Q \to \infty$ gives (2.210) with $K = \lambda_1/\mu_0$. □

We can finally deduce the lower bound on S.

Lemma 7. *Provided* $S < V$ *and Lemma 6 holds, a lower bound exists on the speed* S, *namely*

$$S \geq \frac{V}{K+1}. \tag{2.218}$$

Proof. If (2.210) holds then (2.208) becomes

$$1 \leq \left[\frac{KS}{V} + \frac{S}{V} - 1\right]\int_{t_P}^{\infty} a\big(S(t-t_P),t\big)\,\mathrm{d}t,$$

and so

$$S\left(\frac{K+1}{V}\right) - 1 \geq \frac{1}{\int_{t_P}^{\infty} a\big(S(t-t_P),t\big)\,\mathrm{d}t}. \tag{2.219}$$

Assuming $a \geq 0$, (2.219) implies

$$S\left(\frac{K+1}{V}\right) - 1 \geq 0,$$

and rearranging this leads to (2.218). □

Hence, if a travelling wave of speed S exists, then

$$\frac{V}{K+1} \leq S < V. \tag{2.220}$$

For the Linear Model $K = \lambda/\mu$ and so (2.220) reduces to $V' \leq S < V$ as expected.

2.8.4 Travelling wave solution of the Linear Model

We now return to the Linear Model and write this in the form

$$a_t + Va_x = -\lambda a + \mu b \tag{2.221}$$

$$b_t = \lambda a - \mu b. \tag{2.222}$$

Suppose travelling wave solutions exist for both a and b. Then

$$a(x,t) = \bar{a}(x - Ut), \qquad b(x,t) = \bar{b}(x - Ut), \tag{2.223}$$

where U is the travelling wave speed. Substituting these expressions into (2.221) and (2.222), and adding the resulting equations to eliminate the source term, gives

$$(U - V)\bar{a}'(z) + U\bar{b}'(z) = 0. \tag{2.224}$$

where $z = x - Ut$. We assume

$$\{\bar{a}; \bar{b}\} \longrightarrow \{a_l, a_r; b_l, b_r\}, \quad \text{as} \quad z \longrightarrow \mp\infty,$$

with $\lambda a_l = \mu b_l$ and $\lambda a_r = \mu b_r$ so that the solution is commensurate with the boundary conditions. We can then integrate (2.224) from $-\infty$ to z and, after some simple manipulation to emininate $\bar{b}$, this gives the following ODE to solve for $\bar{a}$:

$$(U - V)\bar{a}'(z) - \left[\lambda + \mu\frac{U - V}{U}\right](\bar{a} - a_l) = 0. \tag{2.225}$$

This can be used to find U. Assume that

$$\bar{a}'(z) \longrightarrow 0 \quad \text{as} \quad z \longrightarrow \mp\infty.$$

So, the left hand side of (2.225) is automatically satisfied when $a = a_l$. If $a = a_r$ then the following holds:

$$-\left[\lambda + \mu\frac{U - V}{U}\right](a_r - a_l) = 0.$$

Provided $a_r \neq a_l$, this gives $U = V'$. Thus, if the Linear Model has a travelling wave, it actually moves at the reduced speed V'. In this Chapter we have considered the

mathematical analysis of the Linear Model problem and its extension to a general two equation model with a nonlinear source term. We obtained upper and lower bounds on the solution and, using the exact solution as quoted in (Rhee et al. 1986), found an approximate form of the solution when the parameters λ and μ are assumed large. This enabled us to deduce the reduced speed which is an important feature of these types of models. However, we need to understand how the solution behaves for both large and small values of the parameters and so these phenomena were discussed in Section 2.3.

Chapter 3

The box scheme for linear problems

3.1 Introduction and derivation

This thesis concentrates on finding a robust numerical scheme to approximate problems such as the Linear Model described in Chapter 1. In practical numerical problems of this type there will be numerous chemical species travelling through the groundwater together. It is likely that these large systems will have varying speeds of transport since some of the chemical reaction rates will be fast compared to the advection speed. Hence some of these speeds will be close to the speed of the advection whilst others will be greatly reduced for, as we have seen in the Linear Model, when λ and μ are large the chemical pollutant in the groundwater travels at speed V'.

We wish to find a numerical scheme that can be solved simply in an explicit way, marching forward in time, which can cope with these varying speeds. We require it to be stable and accurate and not force us to use an artificially small time step. This situation has been met before in river flow modelling where the standard scheme is *the box scheme.* This is also called *Preissman four-point scheme* and is based on integral relationships. The derivation is described in detail in (Cunge & Holly Jr 1980, page 65) and we summarise it below.

In this Chapter we study how effective the box scheme is for the Linear Model. We begin by deriving the box scheme for a single linear advection equation with a linear source term and then describe its features. Considerable insight can be obtained by examining the linear advection equation and so we begin by investigating some basic numerical properties of this simple example including the truncation error, stability, the exact solution of the discretised equations and group velocity analysis. Then the box scheme is applied to the Linear Model and we will use both a Fourier and energy

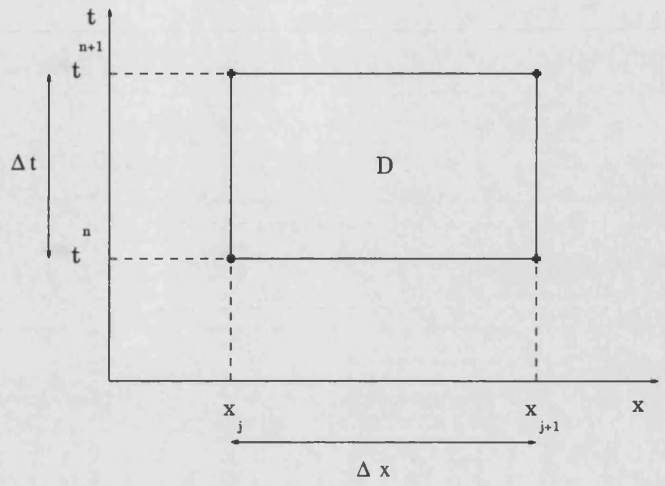

Figure 3-1: *The box D in the (x,t)-plane.*

analysis to investigate stability. We actually discretise the reaction equation using the trapezoidal scheme as it has no spatial derivative and we do not want to introduce extra averaging and therefore potentially more oscillations. The box scheme is well known for producing oscillations when the boundary data is not smooth (as discussed in (Morton & Mayers 1994, page 111)), and we will show that these can be reduced by using a time-weighting of the spatial differences.

A final tool we use to give a great insight into the behaviour of the box scheme is a modified equation analysis. This represents a sequence of PDEs that describes the solution of the discretisation and we can more easily understand the qualitative behaviour of a PDE rather than a system of difference equations. We find modified equation expansions of the box scheme applied to a simple linear advection equation and the trapezoidal scheme to a linear ODE before obtaining the expansions for the Linear Model. We are also able to separate the smooth and oscillatory parts of the solution (which will be justified in detail in Section 3.6.3) and so can obtain modified equation expansions for each. This will enable us to predict where the observed oscillations will occur.

Consider the linear advection equation with linear source term

$$u_t + au_x = bu, \tag{3.1}$$

where $a > 0$ and b are constants, with initial condition $u(x,0) = u^0(x)$ and boundary condition $u(0,t) = f(t)$, given. The box scheme can be derived by integrating (3.1) over the box $D := (x_j, x_{j+1}) \times (t^n, t^{n+1})$ as shown in Figure 3-1. So we consider

$$\iint_D (u_t + au_x)\,\mathrm{d}x\,\mathrm{d}t = b\iint_D u\,\mathrm{d}x\,\mathrm{d}t. \tag{3.2}$$

We can apply the divergence theorem in 2D, as already defined in (2.126) in Chapter 2, to the left hand side of (3.1). Hence (3.2) becomes

$$\oint_{\partial D} \left[-u\,\mathrm{d}x + au\,\mathrm{d}t\right] = b\iint_D u\,\mathrm{d}x\,\mathrm{d}t. \tag{3.3}$$

Since $\mathrm{d}x$ and $\mathrm{d}t$ are both constant along two edges of the box, the integrals in the left hand side of (3.3) can be simplified and we obtain

$$\int_{x_j}^{x_{j+1}} \left[u(x,t_{n+1}) - u(x,t_n)\right]\mathrm{d}x + a\int_{t_n}^{t_{n+1}} \left[u(x_{j+1},t) - u(x_j,t)\right]\mathrm{d}t = b\iint_D u(x,t)\,\mathrm{d}x\,\mathrm{d}t. \tag{3.4}$$

We can now use quadrature rules to approximate these integrals on a general region D, e.g. a rectangular mesh as shown in Figure 3-1. To obtain the box scheme we use the trapezoidal rule; it potentially introduces a spurious mode due to the averaging. This is the notorious chequer-board mode and will be discussed in more detail later in this Chapter.

Instead of deriving the box scheme by using a quadrature rule in (3.4) to approximate the integrals, we can consider a bilinear approximation of u (which we denote as U), based on the values of u defined at the corners of the box D (Forsythe & Wasow 1960, page 332). So

$$\begin{aligned} U(x,t) \;=\; & \frac{1}{\triangle x \triangle t}\left[U_{j+1}^{n+1}(x-x_j)(t-t^n) + U_j^n(x_{j+1}-x)(t^{n+1}-t)\right] \\ & + \frac{1}{\triangle x \triangle t}\left[U_{j+1}^{n}(t^{n+1}-t)(x-x_j) + U_j^{n+1}(x_{j+1}-x)(t-t^n)\right]. \end{aligned} \tag{3.5}$$

If we now suppose (3.4) holds for U then (3.5) can be substituted to evaluate these integrals exactly (which actually gives the trapezoidal rule along the edges of D). The same result is obtained but this derivation is useful to mention as it can be generalised, for example to a quadrilateral mesh.

On a uniform mesh we can also define the box scheme in terms of finite difference operators. Firstly, we describe the notation we will use. Let $0 \le t \le T$ and $0 \le x \le L$, and consider a uniform time step $\triangle t = T/N$ and a uniform spatial step $\triangle x = L/J$, for a given N and J. Let U_j^n denote the numerical approximation of u at $(j\triangle x, n\triangle t)$ for $j = 0, 1, \ldots, J$ and $n = 0, 1, \ldots, N$; i.e. $U_j^n \approx u(x_j, t^n)$ where $x_j := j\triangle x$ and $t^n := n\triangle t$.

The finite difference model box scheme uses averages and differences of the numerical approximation of u at the four nodes of D to approximate these integrals. Define finite difference operators

$$\delta_x U_{j+\frac{1}{2}}^{n+\frac{1}{2}} = U_{j+1}^{n+\frac{1}{2}} - U_j^{n+\frac{1}{2}} \tag{3.6}$$

$$\delta_t U_{j+\frac{1}{2}}^{n+\frac{1}{2}} = U_{j+\frac{1}{2}}^{n+1} - U_{j+\frac{1}{2}}^{n} \tag{3.7}$$

$$\mu_x U_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \tfrac{1}{2}\left(U_{j+1}^{n+\frac{1}{2}} + U_j^{n+\frac{1}{2}}\right) \tag{3.8}$$

$$\mu_t U_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \tfrac{1}{2}\left(U_{j+\frac{1}{2}}^{n+1} + U_{j+\frac{1}{2}}^{n}\right). \tag{3.9}$$

Then the box scheme is given by

$$\frac{\mu_x \delta_t}{\Delta t} U_{j+\frac{1}{2}}^{n+\frac{1}{2}} + a\frac{\mu_t \delta_x}{\Delta x} U_{j+\frac{1}{2}}^{n+\frac{1}{2}} = b\mu_x \mu_t U_{j+\frac{1}{2}}^{n+\frac{1}{2}}, \tag{3.10}$$

which comes directly from discretising (3.1). This can be written as

$$\begin{aligned} \tfrac{1}{2\Delta t}\left(\left[U_{j+1}^{n+1} + U_j^{n+1}\right] - \left[U_{j+1}^{n} + U_j^{n}\right]\right) + \tfrac{a}{2\Delta x}\left(\left[U_{j+1}^{n+1} + U_{j+1}^{n}\right] - \left[U_j^{n+1} + U_j^{n}\right]\right) \\ = \tfrac{1}{4} b\left[U_{j+1}^{n+1} + U_{j+1}^{n} + U_j^{n+1} + U_j^{n}\right]. \end{aligned} \tag{3.11}$$

It is a very compact scheme which uses four neighbouring values of U. It is actually *implicit* as it involves two points at the new time level, but for a linear model with appropriate boundary conditions the solution can be marched away from the boundary and so this involves no extra computation.

If we set $b = 0$ then (3.1) is simply the linear advection equation and (3.11) becomes

$$U_{j+1}^{n+1} = U_j^n + \left(\frac{1-p}{1+p}\right)\left(U_{j+1}^n - U_j^{n+1}\right), \tag{3.12}$$

for $j = 0, \ldots, J-1$ and $n = 0, \ldots, N-1$, where $p := a\Delta t/\Delta x$ is the CFL number. This can be solved sequentially from left to right since we have data prescribed on the left boundary. Note that if the CFL number equals 1 then (3.12) reduces to $U_{j+1}^{n+1} = U_j^n$ and so the box scheme solves the linear equation exactly.

3.2 The linear advection equation

3.2.1 Basic numerical properties

In this section we summarise the traditional numerical tools for analysing finite difference schemes and apply these techniques to the box scheme. The local truncation error is a measure of how well the finite difference equation models the differential equation locally. It is defined by replacing the approximate solution U_j^n in the finite

difference equation by mesh values of the true solution $u(x_j, t^n)$. We now find the local truncation error of the box scheme applied to the conservation law $u_t + au_x = 0$ where a is a positive constant. The finite difference equation is given by (3.12), or, in terms of finite difference operators, by (3.10) with $b = 0$. If we assume smooth solutions then all the terms can be expanded in a Taylor series about the central point $(x_{j+\frac{1}{2}}, t^{n+\frac{1}{2}})$. Since

$$\mu_x \delta_t u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \left[\Delta t u_t + \tfrac{1}{24}\Delta t^3 u_{ttt} + \tfrac{1}{8}\Delta t \Delta x^2 u_{txx} + \ldots\right]_{j+\frac{1}{2}}^{n+\frac{1}{2}} \tag{3.13}$$

$$\mu_t \delta_x u_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \left[\Delta x u_x + \tfrac{1}{8}\Delta x \Delta t^2 u_{ttx} + \tfrac{1}{24}\Delta x^3 u_{txx} + \ldots\right]_{j+\frac{1}{2}}^{n+\frac{1}{2}}, \tag{3.14}$$

the local truncation error becomes

$$\begin{aligned} T_{j+\frac{1}{2}}^{n+\frac{1}{2}} &:= \frac{\mu_x \delta_t u_{j+\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta t} + a\frac{\mu_t \delta_x u_{j+\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x} \\ &= [u_t + a\,u_x] + \tfrac{1}{24}\Delta t^2(u_{ttt} + 3a\,u_{ttx}) + \tfrac{1}{24}\Delta x^2(3u_{txx} + a\,u_{xxx}) + \ldots . \end{aligned} \tag{3.15}$$

We can eliminate the first term since u satisfies the differential equation. Also, we can differentiate $u_t = -au_x$ to replace the u_{ttt}, u_{ttx} and u_{txx} terms by expressions involving only u_{xxx}. Then (3.15) becomes

$$T_{j+\frac{1}{2}}^{n+\frac{1}{2}} = -\tfrac{1}{12}a(a^2\Delta t^2 - \Delta x^2)u_{xxx} + \ldots . \tag{3.16}$$

Hence the box scheme is second order accurate in Δx and Δt.

The local truncation error, which we found above, involves substituting the exact solution into the finite difference scheme. In abstract terms let us consider a partial differential equation in the form

$$Lu = 0, \tag{3.17}$$

where L is the differential operator ($Lu = u_t + au_x$ for the linear advection equation) defined with appropriate initial and boundary conditions. A finite difference scheme can be written as $L_h U = 0$ where U is the numerical solution. Then the local truncation error is defined to be

$$T = L_h(R_h u),$$

and so

$$L_h(U - R_h u) = T,$$

where $R_h u$ is the restriction of u onto the mesh. Hence if we invert L_h we can bound the error between the numerical and true solution. This is classical error analysis.

However, suppose we let

$$0 = L_h U \equiv (L + M)(P_h U), \tag{3.18}$$

where $P_h U$ is the prolongation of U (i.e. the extension so U is now defined for all x and t instead of only at the mesh values). Then

$$L(P_h U - u) = -M P_h U. \tag{3.19}$$

The error is now defined in terms of differential operators. This technique is known as the *modified equation analysis* and will be studied in greater detail in Section 3.5 once we have described the box scheme for the Linear Model. It gives a higher order partial differential equation to describe more accurately the behaviour of the discrete approximation, i.e. the finite difference scheme more accurately approximates the modified equation than then original partial differential equation.

We can also look at stability of the numerical scheme. As pointed out in (Morton & Mayers 1994, pages 110-111), a rigorous Fourier analysis is not valid since our domain is not the whole real line and we do not have periodic boundary conditions. However, it is necessary to consider the substitution of a Fourier mode into the finite difference scheme to consider its possible damping (i.e. substituting $U_j^n = \lambda^n e^{ik(j\triangle x)}$ into (3.12)). It is easy to deduce that

$$|\lambda(k)| = 1. \tag{3.20}$$

From this we can regard the box scheme applied to the linear advection equation as unconditionally stable, provided the equations are solved in the correct direction (i.e. from left to right since $a > 0$ and we have boundary data prescribed at $x = 0$). The condition (3.20) shows there is no damping of the modes and so we might expect there to be oscillations in the numerical solution (unless p is chosen to equal 1). Figure 3-2 shows the box scheme with zero boundary data and a discontinuous square pulse as the initial condition (the dashed line) at two times $t = 0.2$ and $t = 0.5$ for three values of p. In the top two plots $p = 0.25$ and the middle two plots $p = 0.5$; we see that the oscillations increase in number as time progresses but not in size. This supports the theory that the scheme is unconditionally stable. Also, as $\triangle t$ decreases the oscillations increase in number, which is to be expected as we are moving further away from $p = 1$ where we know the numerical solution is exact (and therefore free from oscillations). The bottom two plots show $p > 1$ and we see that the oscillations propagate in the opposite direction. The reason for this will be explained in the next Section (when analysing the group velocity of the box scheme).

As mentioned earlier in this Chapter, the box scheme has a spurious mode due to the averaging: $(-1)^{j+n}$ is always a solution which can be set off by non-smooth boundary data. For the linear advection equation this mode persists for all time and is not

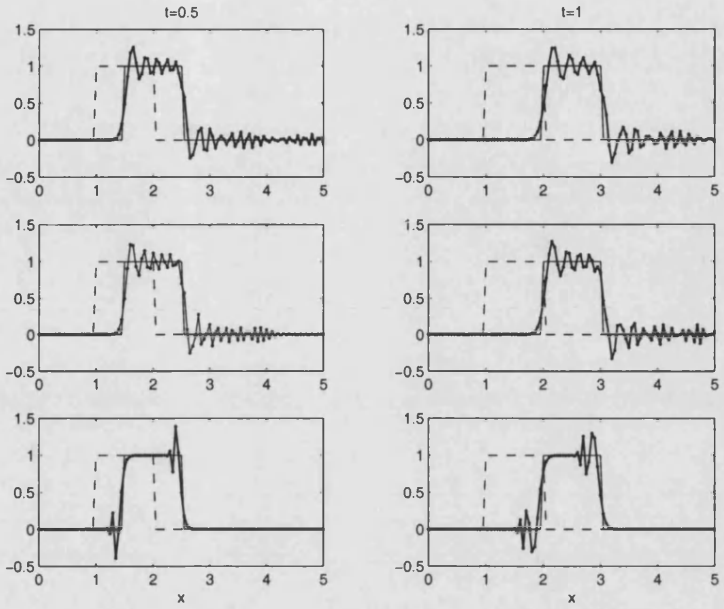

Figure 3-2: *The box scheme (dots joined by an unbroken line) applied to the linear advection equation at fixed times $t = 0.5$ and $t = 1$ for $a = 1$. In the top two plots $p = 0.25$, in the middle two plots $p = 0.5$ and in the bottom two plots $p = 1.25$. The initial condition is a square pulse which is shown as a dashed line and the exact solution is shown as a thin unbroken line.*

damped.

Firstly, let us briefly discuss how the above observations can be explained using the exact solution of the discretised equations. We follow a method described in (Spiegel 1971, page 186) which finds the exact solution of a similar finite difference equation. Assume initial and boundary conditions are

$$u(x,0) = 0, \qquad u(0,t) = \begin{cases} 1, & t_1 \le t \le t_2 \\ 0. & \text{otherwise} \end{cases} \tag{3.21}$$

and consider the box scheme in the form

$$U_j^n = U_{j-1}^{n-1} + \lambda\left(U_{j-1}^n - U_j^{n-1}\right) \qquad j \ge 1, \quad n \ge 1, \tag{3.22}$$

where $\lambda = (p-1)/(p+1)$. Then the conditions in (3.21) become

$$U_j^0 = 0, \qquad U_0^n = \begin{cases} 1, & t_1 \le t^n \le t_2 \\ 0, & \text{otherwise} \end{cases} \tag{3.23}$$

for $n > 0$ and $j \ge 0$. We observe that $|\lambda| < 1$ and so waves are translated along the diagonal at the mesh speed (independent of the real speed) with oscillations away from

it which are trailing either side. Hence, as already discussed, we wish to take p as close to 1 as possible so that λ is small. This is a key feature of the box scheme; although it is unconditionally stable there are oscillations which are dependent on the size of λ.

Suppose we now define difference operators E_1 and E_2 by

$$E_1^{-1}U_j^n = U_{j-1}^n, \qquad E_2^{-1}U_j^n = U_j^{n-1}. \tag{3.24}$$

Then (3.22) can be rewritten as

$$U_j^n = \left(E_2^{-1}\right)\left(E_1^{-1}\right)U_j^n + \lambda\left(E_1^{-1} - E_2^{-1}\right)U_j^n,$$

or

$$E_1 U_j^n = \left(\frac{\lambda + E_2^{-1}}{1 + \lambda E_2^{-1}}\right) U_j^n. \tag{3.25}$$

Following the procedure of (Spiegel 1971, page 186) we consider n to be fixed. Then the solution of (3.25) is

$$U_j^n = \left(\frac{\lambda + E_2^{-1}}{1 + \lambda E_2^{-1}}\right)^j C_n, \tag{3.26}$$

where the C_n, for $n \geq 0$, are to be found using the initial and boundary conditions. In Appendix B (Section B.1) we continue with this analysis by finding the coefficients C_n. We investigate how the oscillations behave as the solution moves away from the diagonal by fixing n and j in turn and allowing the other index to become large. The analysis is very technical, even for the box scheme applied to the linear advection equation; however, we are able to demonstrate that the solution is translated along the diagonal at the mesh speed with oscillations of polynomial size (depending on the sign of λ) trailing out either size.

3.2.2 Group velocity

Energy propagation under dispersive partial differential equations is governed by the quantity known as group velocity. By a dispersive equation we mean one that admits plane wave solutions of the form $e^{i(\omega t - kx)}$, but with the property that the speed of propagation of these waves is not independent of k. Even if an equation is non-dispersive, any discrete model which describes it will be dispersive. We will observe this when studying the modified equation expansion in Section 3.5. So, group velocity is very important in gaining insight to the behaviour of numerical models of partial differential equations. As described in (Trefethen 1982) we can derive the group velocity in one space dimension by a stationary phase argument (due to Lord Kelvin). Suppose that a scalar, linear partial differential equation with constant coefficients admits solutions

of the form

$$u(x,t) = e^{i(\omega t - kx)}. \tag{3.27}$$

For each real *wave number* $k \in \mathbb{Z}$, assume there is a corresponding real *frequency* ω such that (3.27) is a solution. The relation

$$\omega = \omega(k), \tag{3.28}$$

is called the *dispersion relation* for the differential equation. Now, (3.27) propagates rightward at speed

$$c(k) = \frac{\omega(k)}{k}, \tag{3.29}$$

which is called the *phase speed* and the *group speed* is defined as

$$C(k) = \frac{\mathrm{d}\omega}{\mathrm{d}k}(k). \tag{3.30}$$

Consider the simplest hyperbolic equation ($u_t + au_x = 0$) which is non-dispersive. Its dispersion relation is the linear equation

$$\omega(k) = ak, \tag{3.31}$$

and so $c(k) \equiv C(k) \equiv a$. We can now find a similar relation for the box scheme applied to this equation. In discrete form (3.27) becomes

$$U_j^n = e^{i(\omega n \Delta t - kj\Delta x)}. \tag{3.32}$$

Substituting this into (3.12) leads to the dispersion relation

$$\omega(k) = \frac{2}{\Delta t} \tan^{-1}\left[p \tan\left(\frac{k\Delta x}{2}\right)\right], \tag{3.33}$$

where p is the CFL number. The main difference between this relation and (3.31) is that, because the grid is discrete, $\omega(k)$ is multiple-valued and 2π-periodic in $k\Delta x$ and $\omega\Delta t$. Hence it is enough to restrict the domain to $(k,\omega) \in [-\pi/\Delta x, \pi/\Delta x] \times [-\pi/\Delta t, \pi/\Delta t]$. Another thing to note is that the relation in (3.33) is dispersive since neither ω/k or $\mathrm{d}\omega/\mathrm{d}k$ are independent of k. Near $(k,\omega) = (0,0)$ we have that $\omega(k) \approx ak$, but away from the origin this is not the case. The phase speed is given by

$$c(k) = \frac{2}{k\Delta t} \tan^{-1}\left[p \tan\left(\frac{k\Delta x}{2}\right)\right] = a + a\left(\frac{1-p^2}{12}\right)(k\Delta x)^2 + O\big((k\Delta x)^4\big), \tag{3.34}$$

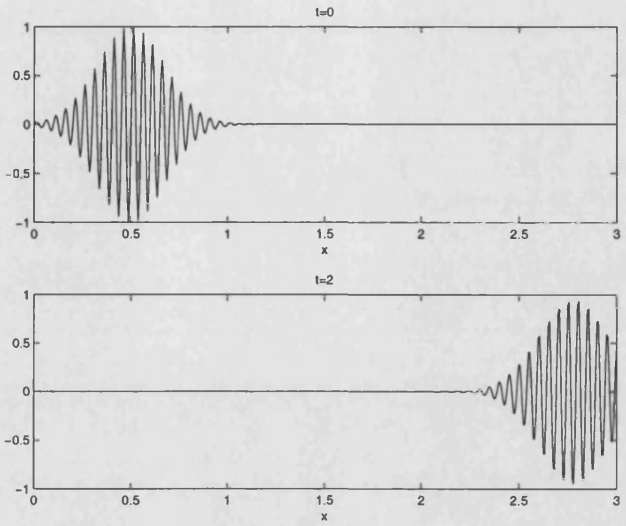

Figure 3-3: *Propagation of a wave packet with* $k\triangle x = 2\pi/8$ *modelled by the box scheme with* $p = 0.4$. *The wave packet does not move at the ideal speed* 1 *but at the group speed* $C \approx 1.13$.

and the group velocity is

$$C(k) := \frac{a}{\cos^2(k\triangle x/2) + p^2 \sin^2(k\triangle x/2)} = a + a\left(\frac{1-p^2}{4}\right)(k\triangle x)^2 + O\big((k\triangle x)^4\big). \tag{3.35}$$

Equation (3.34) really shows why the waves are propagated in different directions depending on whether $p > 1$ or $p < 1$. Also, on comparing (3.34) with (3.35), we see a three fold difference between the $(k\triangle x)^2$ terms. It is precisely $C(k)$ that describes what happens (and has the more significant $(k\triangle x)^2$ term) which we will now demonstrate with two numerical examples.

We use the same two initial conditions as (Trefethen 1982) for $u_t + u_x = 0$, namely a wave packet and a smooth pulse. Consider the following initial wave packet (i.e. a sine wave modulated by a Gaussian pulse):

$$u(x,0) = e^{-16(x-\frac{1}{2})^2} \sin(kx). \tag{3.36}$$

Suppose that $\triangle x = \frac{1}{160}$ and the x-domain is $[0,3]$. If k is chosen so that $k\triangle x = 2\pi/8$ and we set $\triangle t = \frac{1}{400}$ then $p = 0.4$ and Figure 3-3 shows the initial packet (top) and then the box scheme applied to the conservation law after the packet has propagated to $t = 2$ (bottom). The exact solution should move right at speed 1. However, instead of having reached $x = 2.5$, the packet is centred at $x \approx 2.77$, having travelled at speed approximately 1.135. This is precisely the group velocity (and not the phase speed). From (3.34) the phase speed is $c(k) \approx 1.045$ whereas from (3.35) the group velocity is

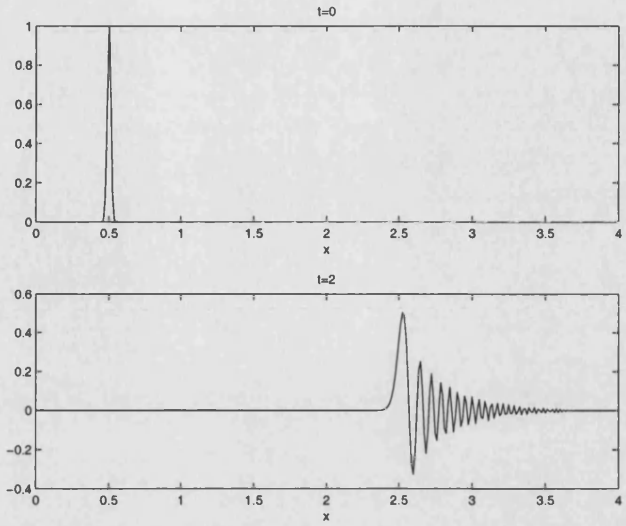

Figure 3-4: *Propagation and dispersion of a narrow pulse modelled by the box scheme with* $p = 0.8$.

$C(k) \approx 1.140$. This analysis demonstrates that there is more to the inaccuracy of a finite difference scheme than truncation error. The wave in the bottom plot of Figure 3-3 differs from the correct solution pointwise but, in fact, it is qualitatively correct. This illustrates why Fourier analysis is a very effective means of analysis.

We now study a Gaussian pulse given by

$$u(x,0) = e^{-3200(x-\frac{1}{2})^2}. \tag{3.37}$$

For this example we take $\triangle x = \frac{1}{80}$, $\triangle t = \frac{1}{100}$ and let the x-domain be $[0,4]$ (again we are using $a = 1$ and so $p = 0.8$). Figure 3-4 shows the initial pulse (top) and then the result of the box scheme being applied to the conservation law after the pulse has propagated to $t = 2$ (bottom). We see that there are oscillations which move faster than the main pulse. As discussed in (Mackenzie 1998) for a similar example we can use the group velocity to predict these observed oscillations. The maximum predicted group velocity is $C(k) = 1.5625$ when $k = \pm\pi/\triangle x$. This approximately matches the velocity at which the fastest oscillations are observed to move. For the low frequency modes where $k\triangle x$ is close to zero, we find that $C(k) = a$ which accounts for the main pulse moving at approximately the correct phase speed. Note that for $p > 1$ the higher wave numbers have lower group speeds and so would lag behind the main pulse.

3.3 The box scheme applied to the Linear Model

We now apply the box scheme to the Linear Model written as a conservation equation coupled with a reaction equation, i.e.

$$a_t + b_t + V a_x = 0 \tag{3.38}$$

$$b_t = \lambda a - \mu b. \tag{3.39}$$

Let A_j^n and B_j^n denote the numerical approximation of a and b respectively at $(j\triangle x, n\triangle t)$ for $j = 0,\ldots,J$ and $n = 0,\ldots,N$. In terms of finite difference operators (3.38) and (3.39) become

$$\mu_x\delta_t A_{j+\frac{1}{2}}^{n+\frac{1}{2}} + \mu_x\delta_t B_{j+\frac{1}{2}}^{n+\frac{1}{2}} + p\mu_t\delta_x A_{j+\frac{1}{2}}^{n+\frac{1}{2}} = 0, \tag{3.40}$$

and

$$\mu_x\delta_t B_{j+\frac{1}{2}}^{n+\frac{1}{2}} = \mu_x\mu_t(\lambda' A_{j+\frac{1}{2}}^{n+\frac{1}{2}} - \mu' B_{j+\frac{1}{2}}^{n+\frac{1}{2}}), \tag{3.41}$$

with p denoting the CFL number $p = V\triangle t/\triangle x$ and $\lambda' := \lambda\triangle t$ and $\mu' := \mu\triangle t$. However (3.41) contains an unnecessary μ_x averaging, which requires an unnecessary boundary condition. Hence we eliminate it and replace (3.41) by

$$\delta_t B_{j+1}^{n+\frac{1}{2}} = \mu_t(\lambda' A_{j+1}^{n+\frac{1}{2}} - \mu' B_{j+1}^{n+\frac{1}{2}}), \tag{3.42}$$

which is simply the trapezoidal scheme in time. As for the box scheme applied to the linear advection equation, the box scheme applied to both equations in the Linear Model (given by (3.40) and (3.41)) has an extra spurious mode. Again $(-1)^{j+n}$ is always a solution which can be set off by non-smooth boundary data. This is very bad when the box scheme is applied to the reaction equation as in (3.41), but much less worse for the trapezoidal scheme (as in (3.42)) since there is no averaging. We will see this in Section 3.6.3 when we will separate the smooth numerical solution and the spurious oscillatory numerical solution and then use the modified equation analysis to describe these different phenomena.

Equations (3.40) and (3.42) can be re-written as

$$\begin{aligned}(1+p)A_{j+1}^{n+1} + (1-p)A_j^{n+1} - (1-p)A_{j+1}^n - (1+p)A_j^n \\ +(B_{j+1}^{n+1} - B_{j+1}^n) + (B_j^{n+1} - B_j^n) = 0\end{aligned} \tag{3.43}$$

$$B_{j+1}^{n+1} = \frac{\frac{1}{2}\lambda'}{1+\frac{1}{2}\mu'}(A_{j+1}^{n+1} + A_{j+1}^n) + \left(\frac{1-\frac{1}{2}\mu'}{1+\frac{1}{2}\mu'}\right) B_{j+1}^n. \tag{3.44}$$

This is an explicit scheme when an initial condition and a one sided boundary condition are imposed. From (3.44), B_{j+1}^{n+1} can be substituted into (3.43) to obtain an explicit formula to solve for A_{j+1}^{n+1}. Once this is known B_{j+1}^{n+1} can easily be found using (3.44).

A straightforward calculation shows that this scheme is second order accurate for all Δx and Δt by finding the local truncation error: because the box scheme is centred in x and t and the trapezoidal scheme is centred in t, all the odd order terms cancel. A Fourier analysis of this method shows the scheme is Lax-Richtymer stable for all Δx and Δt. The details are quite technical and so the proof is given in Appendix B (Section B.3). Note that, before this proof, Appendix B (Section B.2) considers a simpler finite difference scheme for solving the Linear Model (3.38) and (3.39). It uses an explicit scheme to discretise the conservation equation and an implicit scheme to discretise the reaction equation. We call this the ETIR method (explicit transport and implicit reaction) and find a necessary condition for the scheme to be Lax-Richtymer stable. This condition implies we are limited as to how large we can choose the time step Δt. Hence we do not consider the ETIR method in any more detail.

From now on we refer to (3.43) and (3.44) as the *box-trap scheme.* The Fourier analysis in Appendix B only gives a necessary condition for stability. We now consider an energy analysis to investigate sufficient stability conditions for the box-trap scheme.

3.3.1 Energy analysis

Let us briefly return to considering the partial differential equations describing the Linear Model. These can be written in the form

$$a_t + V a_x = -\lambda a + \mu b \tag{3.45}$$

$$b_t = \lambda a - \mu b. \tag{3.46}$$

We multiply (3.45) by a, (3.46) by b and then integrate the resulting equations with respect to x over $(-\infty, \infty)$. This gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{\infty} \tfrac{1}{2}a^2\,\mathrm{d}x + V\int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}x}(\tfrac{1}{2}a^2)\,\mathrm{d}x = -\int_{-\infty}^{\infty} a(\lambda a - \mu b)\,\mathrm{d}x \tag{3.47}$$

$$\frac{d}{dt}\int_{-\infty}^{\infty} \tfrac{1}{2}b^2\,\mathrm{d}x = \int_{-\infty}^{\infty} b(\lambda a - \mu b)\,\mathrm{d}x. \tag{3.48}$$

Now multiplying (3.47) and (3.48) by λ and μ respectively and adding the resulting equations leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{\infty} \tfrac{1}{2}\left(\lambda a^2 + \mu b^2\right)\,\mathrm{d}x + \tfrac{1}{2}\lambda V\left\{a^2\big|_{\infty} - a^2\big|_{-\infty}\right\} = -\int_{-\infty}^{\infty} \left(\lambda^2 a^2 - 2\mu\lambda ab + \mu^2 b^2\right)\,\mathrm{d}x.$$

Assuming that $a(x,t) \to 0$ as $x \to \pm\infty$ for all t we can simplify the above expression

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{\infty} \tfrac{1}{2}\left(\lambda a^2 + \mu b^2\right)\,\mathrm{d}x = -\int_{-\infty}^{\infty} (\lambda a - \mu b)^2\,\mathrm{d}x \leq 0. \tag{3.49}$$

This shows that, in the l^2 norm a and b are decreasing, and therefore bounded by their initial values. We would like to obtain a similar result for the discrete form (i.e. the box-trap scheme applied to these equations). Consider the box-trap scheme written in terms of finite difference operators as in (3.40) and (3.42). We multiply (3.42) by μ_x and set

$$S := \lambda A - \mu B. \tag{3.50}$$

This leads to

$$\mu_x\left[A^{n+1}_{j+\frac{1}{2}} - A^n_{j+\frac{1}{2}}\right] + \tfrac{1}{2}p\delta_x\left[A^{n+1}_{j+\frac{1}{2}} + A^n_{j+\frac{1}{2}}\right] = -\tfrac{1}{2}\triangle t\mu_x\left[S^{n+1}_{j+\frac{1}{2}} + S^n_{j+\frac{1}{2}}\right] \tag{3.51}$$

$$\mu_x\left[B^{n+1}_{j+\frac{1}{2}} - B^n_{j+\frac{1}{2}}\right] = \tfrac{1}{2}\triangle t\mu_x\left[S^{n+1}_{j+\frac{1}{2}} + S^n_{j+\frac{1}{2}}\right]. \tag{3.52}$$

For convenience define

$$\bar{A}^n := \mu_x A^n_{j+\frac{1}{2}}, \quad \bar{B}^n := \mu_x B^n_{j+\frac{1}{2}} \quad \bar{S}^n := \mu_x S^n_{j+\frac{1}{2}}, \tag{3.53}$$

and the l^2 norm for $\bar{A}^n$

$$\|\bar{A}^n\|_2 = \left\{\triangle x \sum_{j=0}^{J-1} \left(\mu_x A^n_{j+\frac{1}{2}}\right)^2\right\}^{\frac{1}{2}}. \tag{3.54}$$

Then the inner product is given by

$$\langle \bar{A}^n, \bar{B}^n\rangle_2 = \triangle x \sum_{j=0}^{J-1} \left(\mu_x A^n_{j+\frac{1}{2}}\right)\left(\mu_x B^n_{j+\frac{1}{2}}\right), \tag{3.55}$$

and so $\|\bar{A}^n\|_2 = \langle \bar{A}^n, \bar{A}^n\rangle_2^{\frac{1}{2}}$. Analogous to the procedure carried out in the continuous case we wish to investigate the sum of a particular combination of A and B over j. Both sides of (3.51) and (3.52) are multiplied by $\bar{A}^{n+1} + \bar{A}^n$ and $\bar{B}^{n+1} + \bar{B}^n$ respectively. We also use the summation by parts result

$$\sum_{j=0}^{J-1} \mu_x\left[A^{n+1}_{j+\frac{1}{2}} + A^n_{j+\frac{1}{2}}\right]\delta_x\left[A^{n+1}_{j+\frac{1}{2}} + A^n_{j+\frac{1}{2}}\right] = \left[A^{n+1}_J + A^n_J\right]^2 - \left[A^{n+1}_0 + A^n_0\right]^2. \tag{3.56}$$

Assume the boundary condition is a short injection of some chemical pollutant. This means that after a finite time $A^n_0 = 0$ for all n. We restrict attention to this case. Summing over j (and multiplying by $\triangle x$) gives

$$\begin{aligned} \|\bar{A}^{n+1}\|^2 - \|\bar{A}^n\|^2 + \tfrac{1}{2}V\triangle t\left[A^{n+1}_J + A^n_J\right]^2 &= -\tfrac{1}{2}\triangle t\langle \bar{A}^{n+1} + \bar{A}^n, \bar{S}^{n+1} + \bar{S}^n\rangle \\ \|\bar{B}^{n+1}\|^2 - \|\bar{B}^n\|^2 &= \tfrac{1}{2}\triangle t\langle \bar{B}^{n+1} + \bar{B}^n, \bar{S}^{n+1} + \bar{S}^n\rangle. \end{aligned}$$

Multiplying the first expression by λ, the second by μ and adding gives

$$\begin{aligned}\lambda\|\bar{A}^{n+1}\|^2+\mu\|\bar{B}^{n+1}\|^2 &= \lambda\|\bar{A}^{n}\|^2+\mu\|\bar{B}^{n}\|^2 \\ &\quad -\tfrac{1}{2}\triangle t\|\bar{S}^{n+1}+\bar{S}^{n}\|^2-\tfrac{1}{2}V\triangle t\left[A_J^{n+1}+A_J^n\right]^2. \quad (3.57)\end{aligned}$$

Hence we have

$$\lambda\|\bar{A}^{n+1}\|^2+\mu\|\bar{B}^{n+1}\|^2 \leq \lambda\|\bar{A}^{n}\|^2+\mu\|\bar{B}^{n}\|^2, \quad (3.58)$$

which implies stability of the averages $\mu_x A$ and $\mu_x B$. However, this does not cover the $(-1)^j$ oscillations and so we have to consider their potential growth separately. We will need to make a link back from the averages to the nodal values. Since we have set the boundary condition to be zero every nodal value can be obtained from the averages by a recurrence ($A_1^n=\bar{A}_{\frac{1}{2}}^n$, $A_2^n=\bar{A}_{\frac{3}{2}}^n-\bar{A}_{\frac{1}{2}}^n$ etc.). Hence the mapping from the averages to the nodal values goes through an oscillatory matrix, i.e. $\mathbf{A}^n=2T\bar{\mathbf{A}}^n$ with

$$T_{ij}=\begin{cases}0, & i<j\\ (-1)^{i+j}, & \text{otherwise.}\end{cases} \quad (3.59)$$

It can be shown that (by examining the l^2 norm of the matrix)

$$\|\mathbf{A}^n\|_2 \leq \sqrt{2J(J+1)}\|\bar{\mathbf{A}}^n\|_2, \quad (3.60)$$

where $\mathbf{A}^n=(A_1^n,\ldots,A_J^n)^T$. Hence there is potential linear growth which is consistent with the well known phenomenon that imposing inappropriate boundary conditions will potentially cause a linear growth in the oscillatory mode (Richtymer & Morton 1967, pages 131-167).

We will be able to damp this potential linear growth by introducing a weighting in the time averaging for the spatial derivative: this will be discussed in greater detail in Section 4.3.3. Before discussing this modification, we show some numerical results of the box-trap scheme.

3.3.2 Numerical results

Figures 3-5 and 3-6 show some numerical results for the box-trap scheme: the concentration a is given for both small and large values of λ and μ and for two sets of boundary data (a discontinuous square pulse in the top plot and a smooth Gaussian pulse in the bottom). We expect there to be more oscillations for the square pulse as the solution will not be sufficiently smooth. Physically, the solution represents the injection of a short pulse of chemical pollutant at time $t=0$ which will move through the ground and diffuse as time increases. In the remainder of this Thesis we present the results in the following way: the natural thing to do with our initial and boundary conditions

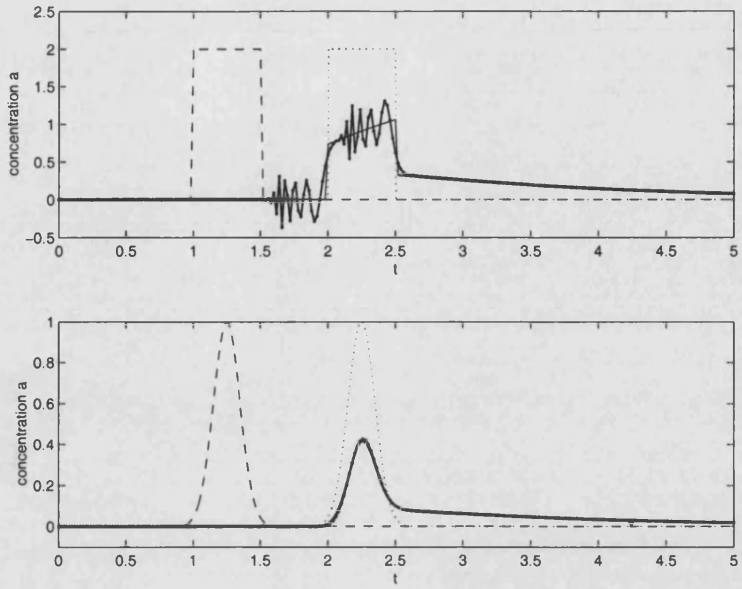

Figure 3-5: *Solution a at fixed $x = 1$: the dashed line indicates the boundary condition (a square pulse in the top plot and a Gaussian curve in the bottom), the thin unbroken line indicates the exact solution, the dots joined by an unbroken line indicate the box-trap scheme and the dotted line indicates the linear advection solution. In both cases $\lambda = 1$, $\mu = 1$, $\triangle x = 0.025$ and $\triangle t = 0.02$ (and so $p = 0.8$).*

is to fix x and investigate what happens in time, since we wish to observe how the chemical pollutant is spreading out at fixed points in space. We will assume that the problem has been nondimensionalised, and so for all numerical experiments presented (unless otherwise stated), we take $V = 1$ and $x = 1$. We also show the solution of the linear advection equation $a_t + Va_x = 0$ in these figures, plotted as a dotted line. This highlights the fact that, when λ and μ are small, the chemical pollutant moves at the same speed as the advection but as these parameters are increased they move at the reduced speed. In the example shown in Figure 3-6 the reduced speed is $1/10$.

The disadvantage of the box scheme is that the averaging in space and time can generate oscillations in the numerical solution. These are very prominent when the boundary condition for a is not smooth as we can see in the top plots in both Figures 3-5 and 3-6. The oscillations are much worse when λ and μ are small but are still visible as the parameters are increased.

There are some interesting features to observe from Figure 3-6: as well as oscillations, the solution dips below the zero axis just ahead of the wave front. This happens for both sets of boundary data. Figure 3-7 shows a blow up of the solution between $t = 0$ and $t = 5$ (left plots) and $t = 5$ and $t = 10$ (right plots) in these cases. For the square pulse there are two sets of oscillations and the peaks are separated by the width of

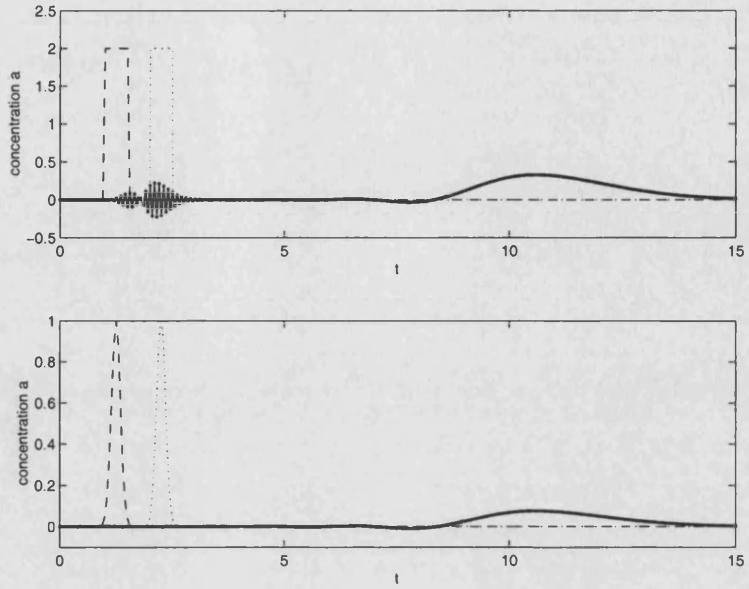

Figure 3-6: *Solution a at fixed $x = 1$: the dashed line indicates the boundary condition (a square pulse in the top plot and a Gaussian curve in the bottom), the thin unbroken line indicates the exact solution (not visible in this case), the dots joined by an unbroken line indicate the box-trap scheme and the dotted line indicates the linear advection solution. In both cases $\lambda = 90$, $\mu = 10$, $\triangle x = 0.0625$ and $\triangle t = 0.05$ (and so $p = 0.8$).*

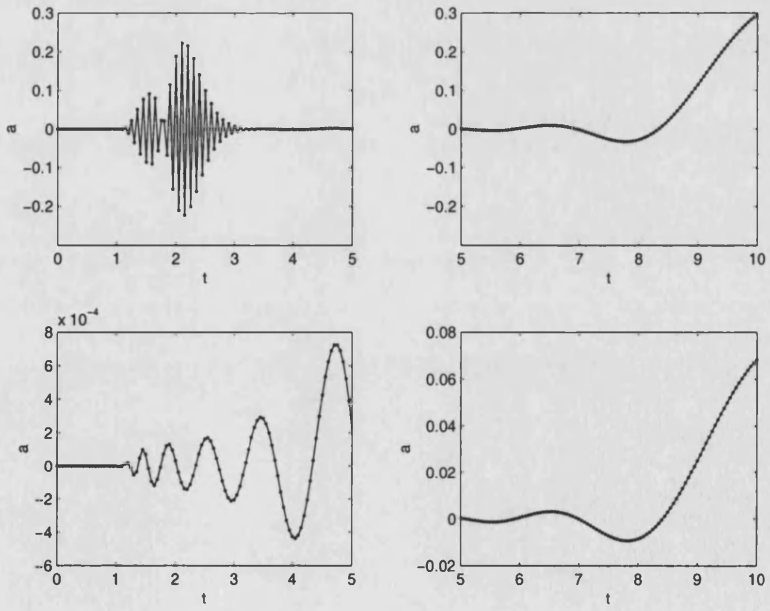

Figure 3-7: *The box-trap scheme is applied to the Linear Model for a square pulse (top two plots) and a Gaussian pulse (bottom two plots) with $\lambda = 90$, $\mu = 10$, $\triangle x = 0.0625$ and $\triangle t = 0.05$ (and so $p = 0.8$). The left two plots show an enlargement of the oscillations and the right two plots shows how the solution becomes negative.*

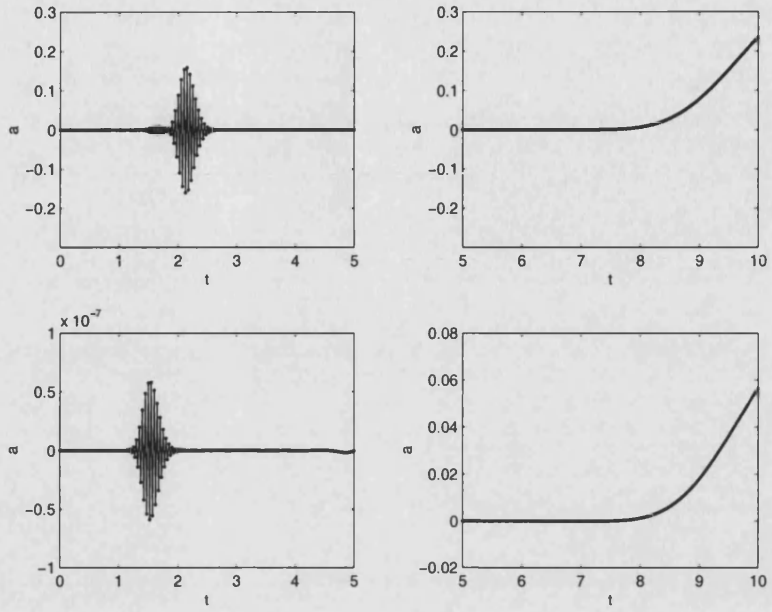

Figure 3-8: *The same situation as shown in Figure 3-7 but with* $\triangle x$ *and* $\triangle t$ *reduced by a factor of two (i.e.* $\triangle x = 0.03125$ *and* $\triangle t = 0.025$ *and so* $p = 0.8$*).*

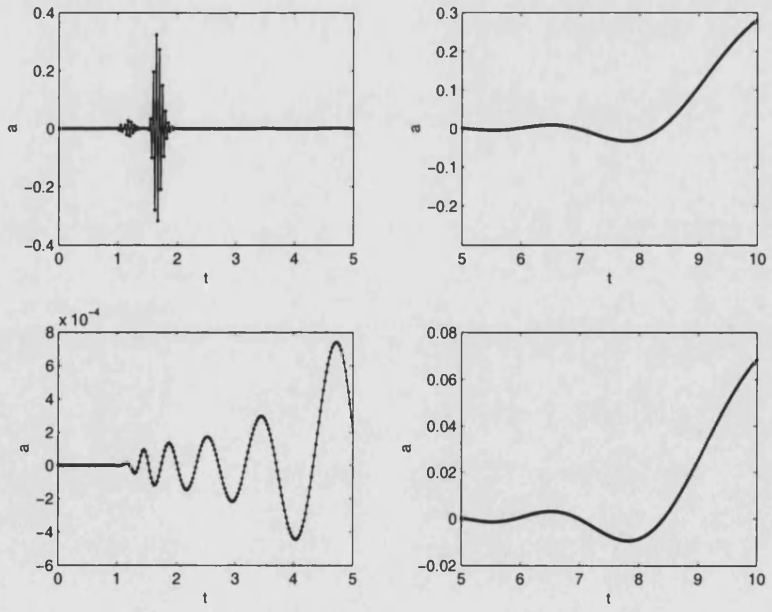

Figure 3-9: *The same situation as shown in Figure 3-7 but with only* $\triangle t$ *reduced by a factor of two (i.e.* $\triangle x = 0.0625$ *and* $\triangle t = 0.025$ *and so* $p = 0.4$*).*

the square pulse. For the Gaussian pulse they do not appear in this form but are still present.

However, in Figure 3-8 we have reduced Δx and Δt by a factor of two. We now see, from the right plots, that the solution does not go negative just ahead of the wave front. Also, the oscillations have reduced in size. There are still two sets of oscillations for the square pulse but these have become much more compact. This is also true for the Gaussian pulse although there is only one set (since the boundary data is smooth). So, for a fixed p, we can eliminate the negativity by reducing Δx and Δt. Unfortunately, this only reduces the oscillations and does not cause them to disappear.

Finally, Figure 3-9 shows a plot of the same situation but with only Δt reduced by a factor of two. In both cases the oscillations have increased; for the Gaussian pulse this is a very small increase since the maximum value of a in the bottom left plot of Figure 3-7 is $\approx 7.245 \times 10^{-4}$ and in Figure 3-9 it is $\approx 7.405 \times 10^{-4}$. Also, the solution is still negative just before the wave front. These Figures illustrate that we need to reduce Δx as well as Δt to increase the accuracy of the solution: decreasing Δt alone will improve the accuracy of the main pulse but will not reduce the oscillations or negativity.

3.4 The weighted box-trap scheme

In this Section we modify the box-trap scheme by using a time-weighting of the spatial differences which is designed to reduce the oscillations. The averaging is now confined to the Transport equation and so we keep the discretisation of the Reaction equation (i.e. (3.42) stays the same), but change (3.40) to become

$$\mu_x \delta_t A_{j+\frac{1}{2}}^{n+\frac{1}{2}} + \mu_x \delta_t B_{j+\frac{1}{2}}^{n+\frac{1}{2}} + p\theta_t \delta_x A_{j+\frac{1}{2}}^{n+\frac{1}{2}} = 0, \tag{3.61}$$

where

$$\theta_t A_j^{n+\frac{1}{2}} = \theta A_j^{n+1} + (1-\theta) A_j^n. \tag{3.62}$$

This can be written as

$$\begin{aligned}(1+2\theta p)A_{j+1}^{n+1} + (1-2\theta p)A_j^{n+1} - [1-2(1-\theta)p]A_{j+1}^n - [1+2(1-\theta)p]A_j^n \\ +\left(B_{j+1}^{n+1} - B_{j+1}^n\right) + \left(B_j^{n+1} - B_j^n\right) = 0.\end{aligned} \tag{3.63}$$

Note that setting $\theta = \frac{1}{2}$ reduces (3.63) to the box scheme. Again a Fourier analysis of this finite difference scheme shows it is Lax-Richtymer stable for all Δx and Δt provided $\theta \geq \frac{1}{2}$ (a proof of which is given in Section B.4 of Appendix B). In the Section below we again consider an energy analysis and will now show that the oscillatory mode can be damped.

Also, the scheme is second order accurate provided $\theta = \frac{1}{2} + O(\Delta t)$. Since only the spatial flux term is modified, it is sufficient to prove this only for the weighted box scheme applied to the linear advection equation $u_t + a u_x = 0$. The discretised equations in this case are simply

$$(1+2\theta p)U_{j+1}^{n+1} + (1-2\theta p)U_j^{n+1} - [1-2(1-\theta)p]U_{j+1}^n - [1+2(1-\theta)p]U_j^n = 0. \quad (3.64)$$

As in Section 2.1 we can expand all the terms in a Taylor series about the central point $(x_{j+\frac{1}{2}}, t^{n+\frac{1}{2}})$ as the origin. Then the truncation error is

$$\begin{aligned} T_{j+\frac{1}{2}}^{n+\frac{1}{2}} &:= \frac{\mu_x \delta_t \left(u_{j+\frac{1}{2}}^{n+\frac{1}{2}}\right)}{\Delta t} + a \frac{\theta_t \delta_x \left(u_{j+\frac{1}{2}}^{n+\frac{1}{2}}\right)}{\Delta x} \\ &= [u_t + a\, u_x] + (\theta - \tfrac{1}{2})\Delta t u_{xt} + \tfrac{1}{24}\Delta t^2 (u_{ttt} + 3a\, u_{ttx}) \\ &+ \tfrac{1}{24}\Delta t^2 (3u_{txx} + a u_{xxx}) + \ldots, \end{aligned} \quad (3.65)$$

which gives second order accuracy in Δx and Δt provided $\theta - \frac{1}{2} = O(\Delta t)$.

Cunge & Holly Jr (1980, page 89) discuss some of the features of the weighted box scheme and the effect of adding the parameter θ. When $\theta = \frac{1}{2}$ the scheme is not dissipative; there is no numerical damping. However, it becomes dissipative by choosing $\frac{1}{2} < \theta \leq 1$. In general it is dispersive; although when the CFL number is equal to 1 the box scheme applied to a linear conservation law is both non-dissipative and non-dispersive.

As we have already observed in this Chapter, when $\theta = \frac{1}{2}$ and the CFL number is not equal to 1, oscillations are observed in the solution. However, when $\theta > \frac{1}{2}$ the scheme becomes diffusive and the oscillations are damped (smeared out). This damping falsifies the amplitude of the solution but often makes the numerical solution more acceptable compared with using $\theta = \frac{1}{2}$.

3.4.1 Energy analysis

Consider the weighted box-trap scheme applied to the Linear Model. For the moment we do not specify how the source term is averaged in the t direction and so we have

$$\mu_x \delta_t A + p\theta_t \delta_x A = -\Delta t \mu_x S^{n+\frac{1}{2}} \quad (3.66)$$

$$\mu_x \delta_t B = \Delta t \mu_x S^{n+\frac{1}{2}}, \quad (3.67)$$

where the discretisation of the reaction equation has again been multiplied by μ_x. This can be expanded to give

$$\mu_x\left[A^{n+1}_{j+\frac{1}{2}} - A^n_{j+\frac{1}{2}}\right] + p\delta_x\left[\theta A^{n+1}_{j+\frac{1}{2}} + (1-\theta)A^n_{j+\frac{1}{2}}\right] = -\triangle t\mu_x S^{n+\frac{1}{2}}_{j+\frac{1}{2}} \tag{3.68}$$

$$\mu_x\left[B^{n+1}_{j+\frac{1}{2}} - B^n_{j+\frac{1}{2}}\right] = \triangle t\mu_x S^{n+\frac{1}{2}}_{j+\frac{1}{2}}. \tag{3.69}$$

Following the same procedure as for the box-trap scheme, we multiply (3.68) and (3.69) by

$$\theta\bar{A}^{n+1} + (1-\theta)\bar{A}^n, \quad \theta\bar{B}^{n+1} + (1-\theta)\bar{B}^n,$$

respectively (where $\bar{A}^n$ etc. are defined in (3.53)), and then sum the resulting equations over j. We now use the following summation by parts result to simplify the second term in (3.68):

$$\sum_{j=0}^{J-1} \mu_x\left[\theta A^{n+1}_{j+\frac{1}{2}} + (1-\theta)A^n_{j+\frac{1}{2}}\right]\delta_x\left[\theta A^{n+1}_{j+\frac{1}{2}} + (1-\theta)A^n_{j+\frac{1}{2}}\right] = \left[\theta A^{n+1}_J + (1-\theta)A^n_J\right]^2 - \left[\theta A^{n+1}_0 + (1-\theta)A^n_0\right]^2.$$

Again assuming $A^n_0 = 0$ for all n past a finite point, we have

$$\theta\|\bar{A}^{n+1}\|^2 + (1-2\theta)\langle\bar{A}^{n+1}, \bar{A}^n\rangle - (1-\theta)\|\bar{A}^n\|^2 + V\triangle t\left[\theta A^{n+1}_J + (1-\theta)A^n_J\right]^2 = -\triangle t\langle\theta\bar{A}^{n+1} + (1-\theta)\bar{A}^n, \bar{S}^{n+\frac{1}{2}}\rangle \tag{3.70}$$

$$\theta\|\bar{B}^{n+1}\|^2 + (1-2\theta)\langle\bar{B}^{n+1}, \bar{B}^n\rangle - (1-\theta)\|\bar{B}^n\|^2 = \triangle t\langle\theta\bar{B}^{n+1} + (1-\theta)\bar{B}^n, \bar{S}^{n+\frac{1}{2}}\rangle. \tag{3.71}$$

Now

$$\langle\bar{A}^{n+1}, \bar{A}^n\rangle \leq \tfrac{1}{2}\left[\|\bar{A}^{n+1}\|^2 + \|\bar{A}^n\|^2\right], \tag{3.72}$$

and so

$$(2\theta-1)\langle\bar{A}^{n+1}, \bar{A}^n\rangle \leq \left(\theta - \tfrac{1}{2}\right)\left[\|\bar{A}^{n+1}\|^2 + \|\bar{A}^n\|^2\right].$$

Hence (3.70) becomes

$$\begin{aligned}\theta\|\bar{A}^{n+1}\|^2 - (1-\theta)\|\bar{A}^n\|^2 \leq\ & \left(\theta - \tfrac{1}{2}\right)\left[\|\bar{A}^{n+1}\|^2 + \|\bar{A}^n\|^2\right] \\ & -\triangle t\langle\theta\bar{A}^{n+1} + (1-\theta)\bar{A}^n, \bar{S}^{n+\frac{1}{2}}\rangle \\ & -V\triangle t\left[\theta A^{n+1}_J + (1-\theta)A^n_J\right]^2.\end{aligned} \tag{3.73}$$

The final term is negative and so we can rearrange to obtain

$$\tfrac{1}{2}\|\bar{A}^{n+1}\|^2 - \tfrac{1}{2}\|\bar{A}^n\|^2 \leq -\triangle t\langle\theta\bar{A}^{n+1} + (1-\theta)\bar{A}^n, \bar{S}^{n+\frac{1}{2}}\rangle. \tag{3.74}$$

Also, using (3.71), we can deduce the same inequality for B, i.e.

$$\tfrac{1}{2}\|\bar{B}^{n+1}\|^2 - \tfrac{1}{2}\|\bar{B}^n\|^2 \leq \Delta t\langle\theta\bar{B}^{n+1} + (1-\theta)\bar{B}^n, \bar{S}^{n+\frac{1}{2}}\rangle. \tag{3.75}$$

We now need to specify how to average the source term in the t direction. If we were to use a weighted average, i.e.

$$\bar{S}^{n+\frac{1}{2}} = \theta\bar{S}^{n+1} + (1-\theta)\bar{S}^n, \tag{3.76}$$

then, multiplying (3.74) by λ, (3.75) by μ and adding, leads to

$$\lambda\|\bar{A}^{n+1}\|^2 + \mu\|\bar{B}^{n+1}\|^2 \leq \lambda\|\bar{A}^n\|^2 + \mu\|\bar{B}^n\|^2 - 2\Delta t\|\theta\bar{S}^{n+1} + (1-\theta)\bar{S}^n\|^2. \tag{3.77}$$

Hence the averages $\mu_x A^n$ and $\mu_x B^n$ are stable. However, because the weighting adds extra diffusion to the system, we only want to replace θ_t by μ_t in the approximation of the spatial derivative (and not the reaction term). So, we set

$$\bar{S}^{n+\frac{1}{2}} = \tfrac{1}{2}(\bar{S}^{n+1} + \bar{S}^n). \tag{3.78}$$

Then, (3.74) and (3.75) become

$$\begin{aligned}\lambda\|\bar{A}^{n+1}\|^2 + \mu\|\bar{B}^{n+1}\|^2 &\leq \lambda\|\bar{A}^n\|^2 + \mu\|\bar{B}^n\|^2 - \Delta t\langle\bar{S}^{n+1} + \bar{S}^n, \theta\bar{S}^{n+1} + (1-\theta)\bar{S}^n\rangle \\ &= \lambda\|\bar{A}^n\|^2 + \mu\|\bar{B}^n\|^2 - \theta\Delta t\|\bar{S}^{n+1}\|^2 - \Delta t\langle\bar{S}^{n+1}, \bar{S}^n\rangle \\ &\quad -(1-\theta)\Delta t\|\bar{S}^n\|^2. \end{aligned} \tag{3.79}$$

Suppose we define $\xi := 2\theta - 1$. Then $0 < \xi \leq 1$ and

$$\theta = \tfrac{1}{2}(1+\xi), \qquad 1-\theta = \tfrac{1}{2}(1-\xi).$$

So (3.79) can be written as

$$\begin{aligned}\lambda\|\bar{A}^{n+1}\|^2 + \mu\|\bar{B}^{n+1}\|^2 + \tfrac{1}{2}\xi\Delta t\|\bar{S}^{n+1}\|^2 &\leq \lambda\|\bar{A}^n\|^2 + \mu\|\bar{B}^n\|^2 + \tfrac{1}{2}\xi\Delta t\|\bar{S}^n\|^2 \\ &\quad -\tfrac{1}{2}\Delta t\left[\|\bar{S}^{n+1}\|^2 + 2\langle\bar{S}^{n+1}, \bar{S}^n\rangle + \|\bar{S}^n\|^2\right] \\ &= \lambda\|\bar{A}^n\|^2 + \mu\|\bar{B}^n\|^2 + \tfrac{1}{2}\xi\Delta t\|\bar{S}^n\|^2 \\ &\quad -\tfrac{1}{2}\Delta t\|\bar{S}^{n+1} + \bar{S}^n\|^2. \end{aligned} \tag{3.80}$$

Finally

$$\lambda\|\bar{A}^{n+1}\|^2 + \mu\|\bar{B}^{n+1}\|^2 + \tfrac{1}{2}\xi\Delta t\|\bar{S}^{n+1}\|^2 \leq \lambda\|\bar{A}^n\|^2 + \mu\|\bar{B}^n\|^2 + \tfrac{1}{2}\xi\Delta t\|\bar{S}^n\|^2, \tag{3.81}$$

and so the averages $\mu_x A^n$ and $\mu_x B^n$ are stable. We still have to consider the $(-1)^j$ oscillatory mode. As discussed in Section 3.3.1 for the box-trap scheme, introducing the weighting θ will damp the potential linear growth of this mode. This can be seen by using the *Godunov-Ryabenkii stability criterion* (Richtymer & Morton 1967, page 152) which, in practice, involves examining all the local normal modes. Suppose we look for normal solution modes of the form

$$A_j^n = \alpha^n \beta^j \hat{A}, \qquad B_j^n = \alpha^n \beta^j \hat{B}. \tag{3.82}$$

We must show that α and β cannot lie outside the unit circle. We only consider the oscillatory mode which corresponds to setting $\beta = -1$ in (3.82). Substituting these solution modes into (3.44) together with (3.63) leads to the following equation which α must satisfy:

$$2p\big(\theta\alpha + (1-\theta)\big)\big[(\alpha - 1) + \tfrac{1}{2}\mu'(\alpha + 1)\big] = 0. \tag{3.83}$$

This has two roots

$$\alpha = -\frac{1-\xi}{1+\xi}, \quad \text{or} \quad \alpha = \frac{1 - \frac{1}{2}\mu'}{1 + \frac{1}{2}\mu'}. \tag{3.84}$$

Thus $|\alpha| < 1$ and so the oscillatory mode is damped for $\theta > \frac{1}{2}$. If $\theta = \frac{1}{2}$ then the first root in (3.84) is 1 and so the oscillatory mode for the box-trap scheme could persist undamped for all time.

3.4.2 Numerical results

We now apply the weighted box-trap scheme, i.e. (3.44) and (3.63), to the Linear Model when the boundary condition is a square pulse to see the reduction in the oscillations. Figures 3-10 and 3-11 show plots for two values of θ (but both with $\theta = \frac{1}{2} + O(\Delta t)$) for small and large values of λ and μ respectively. For both cases of λ and μ we can see that increasing θ from $\frac{1}{2}$ has reduced the oscillations. If we compare the top plot of Figure 3-5 with Figure 3-10 we see that the oscillations keep on reducing as θ increases but that the accuracy gets worse. Hence, although for $\lambda = \mu = 1$ the oscillations have reduced, the weighting introduces a smoothing of the solution.

For large λ and μ the top plots of Figure 3-6 and Figure 3-11 can be compared. We only have to take $\theta = \frac{1}{2} + \Delta t$ for the oscillations to disappear completely. In fact, for large λ and μ, we can actually take θ closer to $\frac{1}{2}$. It would appear that the weighted box-trap scheme removes the oscillations completely and keeps the same accuracy as for the box-trap scheme (this is because there is so much smoothing in the solution anyway). However, for small λ and μ this is not the case: the oscillations are reduced but a numerical dispersion is introduced. We still observe negativity before the wave front for large λ and μ (see Figure 3-11) but, as for the box-trap scheme, this would

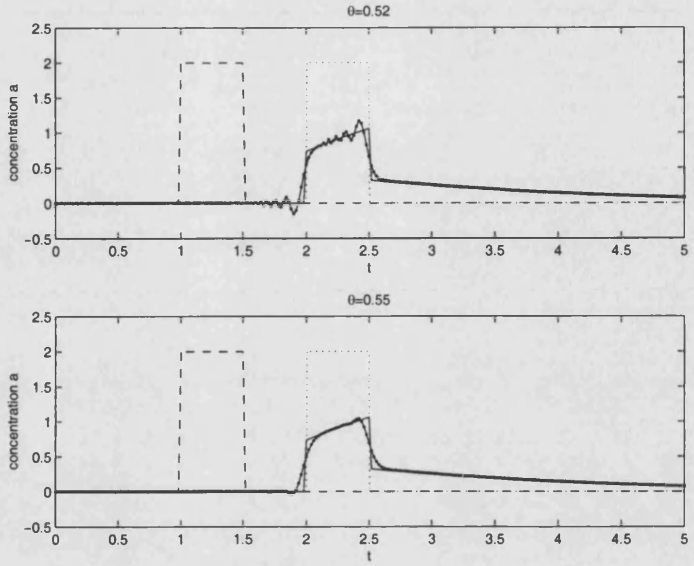

Figure 3-10: *Solution a at fixed $x = 1$: the dashed line indicates the boundary condition, the thin unbroken line indicates the exact solution, the dots joined by an unbroken line indicate the box-trap scheme (in the top plot $\theta = 0.52$ and in the bottom $\theta = 0.55$) and the dotted line indicates the linear advection solution. In both cases $\lambda = 1$, $\mu = 1$, $\triangle x = 0.025$ and $\triangle t = 0.02$ (and so $p = 0.8$).*

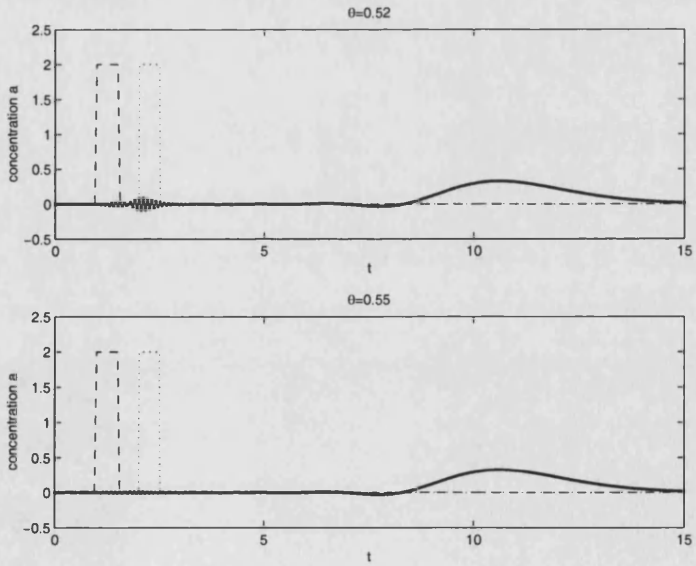

Figure 3-11: *Solution a at fixed $x = 1$: the dashed line indicates the boundary condition, the thin unbroken line indicates the exact solution, the dots joined by an unbroken line indicate the box-trap scheme (in the top plot $\theta = 0.52$ and in the bottom $\theta = 0.55$) and the dotted line indicates the linear advection solution. In both cases $\lambda = 90$, $\mu = 10$, $\triangle x = 0.0625$ and $\triangle t = 0.05$ (and so $p = 0.8$).*

be eliminated by reducing Δx.

We would like to be able to take any size of Δt but for small λ and μ we might have to restrict ourselves so as to avoid the smoothing which occurs when using the weighting: if Δt is too large then θ will be significantly larger than $\frac{1}{2}$ (to keep second order accuracy) and this will cause severe damping.

3.5 Modified equation analysis

Various qualities and properties of a finite difference scheme of a given PDE can be analysed by studying the modified partial differential equation of (3.17) and (3.18), known as *modified equation analysis.* These include order of accuracy, consistency, dissipation and dispersion. Apart from the round-off error, the modified equation analysis represents an asymptotic series of PDEs that initially gives a more accurate representation the behaviour of the solution of the difference scheme. The expansion is derived by first expanding each term of a difference scheme in a Taylor Series and then progressively eliminating time derivatives of higher than first order in favour of higher and higher spatial derivatives. Basically, to study the behaviour of solutions to the difference equations we are modelling the difference equation by a sequence of differential equations. This was seen in Section 3.2 where we showed that the error in the numerical scheme could be defined in terms of differential operators (see (3.19)). Of course the difference equation was originally derived by approximating a PDE, and so the original PDE is a model for the difference equation, but there are other differential equations that are better models. In other words, there are other PDEs that the numerical method solves more accurately than the original PDE. It is often easier to predict the qualitative behaviour of a PDE than of a system of difference equations. Warming & Hyett (1974) describe this method in more detail. At the moment it is the qualitative behaviour of a particular numerical method that we wish to understand.

The modified equation approach is insufficiently powerful to study the stability of finite difference schemes. It involves an expansion in the differential operators which is only valid for small $k\Delta x$ where k is the wave number and Δx is the spatial step length. Basically, when we truncate, the higher order terms might blow up if $k\Delta x$ is large. The mode $k\Delta x = \pi$ is often the most unstable but this cannot be reached with this expansion (because it does not converge). We can obtain information on stability as $k\Delta x \to 0$ but not for the case $k\Delta x \to \pi$. So, the modified equation approach can give necessary but not sufficient conditions for stability.

By the same token it cannot describe the mesh scale oscillations induced by the spurious mode in the box scheme. To do this we have to separate the smooth numerical solution and the spurious oscillatory numerical solution (which will be done in Section 3.6.3).

3.5.1 The linear advection equation

Before analysing the box-trap scheme applied to the Linear Model, we first carry out the modified equation analysis of two simple problems using the same numerical schemes (i.e the box scheme and the trapezoidal scheme) to gain insight into this technique applied to simple linear problems. First consider the linear advection equation $u_t + au_x = 0$ which we studied in some detail in Section 3.2 and where we briefly mentioned how the modified equation expansion can be used as a tool for analysing the oscillations induced by the box scheme. In terms of finite difference operators the box scheme is simply given by (3.10), or rearranging we obtain

$$\frac{\delta_t \mu_t^{-1}}{\Delta t} U + a \frac{\delta_x \mu_x^{-1}}{\Delta x} U = 0, \tag{3.85}$$

where we have dropped the $(\cdot)_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ notation for convenience. We now suppose (3.85) holds for $\tilde{U}$, where $\tilde{U}$ is the prolongation of U. The finite difference operators defined by (3.6)–(3.9) can be expanded in terms of differential operators using Taylor series expansions, i.e.

$$\begin{aligned}
\delta_x &= \Delta x \left[1 + \tfrac{1}{24}\Delta x^2 \partial_x^2 + O(\Delta x^4)\right] \partial_x && (3.86)\\
\delta_t &= \Delta t \left[1 + \tfrac{1}{24}\Delta t^2 \partial_t^2 + O(\Delta t^4)\right] \partial_t && (3.87)\\
\mu_x &= 1 + \tfrac{1}{8}\Delta x^2 \partial_x^2 + O(\Delta x^4) && (3.88)\\
\mu_t &= 1 + \tfrac{1}{8}\Delta t^2 \partial_t^2 + O(\Delta t^4). && (3.89)
\end{aligned}$$

Subsequently, we will use the notation "..." to mean either $O(\Delta x^4)$ or $O(\Delta t^4)$. If we define the operators

$$\mathcal{D}_t := \frac{\delta_t \mu_t^{-1}}{\Delta t}, \quad \mathcal{D}_x := \frac{\delta_x \mu_x^{-1}}{\Delta x}, \tag{3.90}$$

then (3.85) gives a simple expression relating the t-derivatives (from the $\mathcal{D}_t$ term) to the x-derivatives (from the $\mathcal{D}_x$ term)

$$(\mathcal{D}_t + a\mathcal{D}_x)\tilde{U} = 0. \tag{3.91}$$

Substituting the expansions (3.86)–(3.89) into the operators (3.90) leads to expressions for $\mathcal{D}_t$ and $\mathcal{D}_x$ (up to and including Δx^2 and Δt^2 terms)

$$\begin{aligned}
\mathcal{D}_t &= \left(1 - \tfrac{1}{12}\Delta t^2 \partial_t^2 + \ldots\right) \partial_t && (3.92)\\
\mathcal{D}_x &= \left(1 - \tfrac{1}{12}\Delta x^2 \partial_x^2 + \ldots\right) \partial_x. && (3.93)
\end{aligned}$$

Then (3.91) becomes

$$\left(1 - \tfrac{1}{12}\Delta t^2 \partial_t^2 + \ldots\right) \tilde{U}_t = -a \left(1 - \tfrac{1}{12}\Delta x^2 \partial_x^2 + \ldots\right) \tilde{U}_x, \tag{3.94}$$

or

$$\begin{aligned} \tilde{U}_t &= -a \left(1 - \tfrac{1}{12}\Delta t^2 \partial_t^2 + \ldots\right)^{-1} \left(1 - \tfrac{1}{12}\Delta x^2 \partial_x^2 + \ldots\right) \tilde{U}_x, \\ &= -a \left(1 + \tfrac{1}{12}\Delta t^2 \partial_t^2 - \tfrac{1}{12}\Delta x^2 \partial_x^2\right) \tilde{U}_x + \ldots. \end{aligned} \tag{3.95}$$

We now differentiate (3.95) with respect to x and then t to eliminate the ∂_t^2 term. It is easy to show that

$$\partial_t^2 \tilde{U}_x = a^2 \partial_x^2 \tilde{U}_x + \ldots.$$

Substituting this expression back into (3.95) gives the modified equation expansion

$$\tilde{U}_t = -a\tilde{U}_x - \frac{a}{12}\left(a^2\Delta t^2 - \Delta x^2\right) \tilde{U}_{xxx} + \ldots. \tag{3.96}$$

which exactly matches the Truncation error in (3.16). The partial differential equation obtained by truncating this expansion after the first two terms has a dispersion term which depends on the value of the CFL number, $p := \frac{a\Delta t}{\Delta x}$. This analysis supports the theory of Section 3.2.2: if $p = 1$ then this term disappears and we would expect there to be little dispersion (which is a special case for the linear advection equation because for $p = 1$ the difference scheme is exact). Also, as p travels through unity the oscillations change the direction in which they propagate. We observed this in Figure 3-2 and can see this directly from (3.96) as the sign of the dispersion term U_{xxx} changes for $p < 1$ and $p > 1$. This was also shown to be true by examining the group velocity.

3.5.2 A simple ordinary differential equation (ODE)

The second problem to consider is the first order ODE given by

$$b_t = -\mu b, \tag{3.97}$$

where μ is an arbitrary constant. Applying the trapezoidal scheme (in terms of differential operators) to (3.97) gives

$$\frac{\delta_t}{\Delta t} B = -\mu \mu_t B, \tag{3.98}$$

(apologies for the two μ's here) where B is the numerical approximation of b. This can be rewritten as

$$\mathcal{D}_t B = -\mu B, \tag{3.99}$$

where $\mathcal{D}_t$ is given in (3.90). We now apply (3.99) to the prolongation of B (which we denote by $\tilde{B}$) and invert the expansion of $\mathcal{D}_t$ in (3.92). Hence, the modified equation expansion for (3.98) is given by

$$\begin{aligned} \tilde{B}_t &= -\mu\left(1 - \tfrac{1}{12}\Delta t^2\partial_t^2 + \ldots\right)^{-1}\tilde{B} \\ &= -\mu\left(1 + \tfrac{1}{12}\Delta t^2\partial_t^2 + \ldots\right)\tilde{B}. \end{aligned} \tag{3.100}$$

The expansion (3.100) represents a sequence of actual ODEs solved when the numerical scheme is computed. Hence studying this sequence will give us insight into the numerical solution of (3.97). If we truncate the expansion in (3.100) after the second term we obtain the following ODE:

$$\tfrac{1}{12}\mu\Delta t^2 B_{tt} + B_t + \mu B = 0, \tag{3.101}$$

dropping the $\tilde{(\cdot)}$ notation for convenience. The auxiliary equation is given by

$$\tfrac{1}{12}\mu\Delta t^2 p^2 + p + \mu = 0,$$

which has two roots

$$p_\pm = \frac{6}{\mu\Delta t^2}\left[-1 \pm \sqrt{\left(1 - \tfrac{1}{3}\mu^2\Delta t^2\right)}\right], \tag{3.102}$$

and so the solution of (3.101) is

$$B(t) = C_1 e^{p_+ t} + C_2 e^{p_- t}. \tag{3.103}$$

We now make some key points about this analysis. The roots in (3.102) show that the numerical scheme is second order accurate. Also, if $\mu\Delta t > \sqrt{3}$, the roots are complex and so the solution of (3.101) is oscillatory; however, if $\mu\Delta t < \sqrt{3}$, the solution decays exponentially. The spurious mode $e^{p_- t}$ will decay very quickly to zero and so the positive root will be the main contribution (which also decays). We can see this by expanding the term under the square root. So, provided $|\mu^2\Delta t^2| < 3$ we have

$$p_+ = -\mu\left(1 + \tfrac{1}{12}\mu^2\Delta t^2 + \tfrac{1}{72}\mu^4\Delta t^4 + \ldots\right), \tag{3.104}$$

and

$$p_- = -\frac{1}{\mu\Delta t^2}\left[12 - \mu^2\Delta t^2\right] + \tfrac{1}{12}\mu^3\Delta t^2 + \tfrac{1}{72}\mu^5\Delta t^4 + \ldots. \tag{3.105}$$

The term in the square brackets in (3.105) is negative (assuming $\mu > 0$) and so this root will decay very quickly. The expansion (3.104) shows that the positive root behaves like $e^{-\mu t}$, which is to be expected.

We could have solved (3.98) as a finite difference scheme directly. Then

$$B^{n+1} = \left(\frac{1-\frac{1}{2}\mu\Delta t}{1+\frac{1}{2}\mu\Delta t}\right) B^n. \tag{3.106}$$

Note that the expansion of the term in brackets in (3.106), for small $\mu\Delta t$, is given by

$$\begin{aligned}\left(\frac{1-\frac{1}{2}\mu\Delta t}{1+\frac{1}{2}\mu\Delta t}\right) &= (1-\tfrac{1}{2}\mu\Delta t)\,(1-\tfrac{1}{2}\mu\Delta t+\tfrac{1}{4}\mu^2\Delta t^2-\tfrac{1}{8}\mu^3\Delta t^3-\ldots)\\ &= 1-\mu\Delta t+\tfrac{1}{2}\mu^2\Delta t^2-\tfrac{1}{4}\mu^3\Delta t^3+O(\Delta t^4). \end{aligned}\tag{3.107}$$

A simple calculation shows that $e^{p+\Delta t}$ tends to the expansion (3.107).

In conclusion we know, from studying the finite difference scheme directly, that there are oscillations if $\mu\Delta t > 2$. However, in a large system we also wish to apply the modified equation analysis to ODEs and so would like to understand (3.100). The modified equation analysis does predict these oscillations but the condition was found to be $\mu\Delta t > \sqrt{3}$. The expansion in (3.107) is exact whereas we truncated the modified equation expansion after two terms. This means that the condition predicted from the modified equation expansion has an error proportional in magnitude to the first truncated term. On the other hand, it does give the same expansion that is found when analysing the ODE in terms of the finite difference scheme and so suggests it will be a useful tool to help explain the oscillations that occur in the Linear Model.

Figure 3-12 shows plots of the solution b with fixed $b(0) = 0.5$ and $\mu = 20$ for various $\mu\Delta t$'s. These are given by $\mu\Delta t = 0.5$ (top left plot), $\mu\Delta t = 1.25$ (top right plot), $\mu\Delta t = 2.5$ (bottom left plot) and $\mu\Delta t = 5$ (bottom right plot). These confirm that the numerical scheme oscillates when $\mu\Delta t > 2$. This condition arises when μ is relatively large. In the Linear Model we want to be able to take a large range of values for the parameters λ and μ and so these oscillations could be a problem.

3.6 Modified equation analysis of the box-trap scheme

We now find the modified equation expansion of the box-trap scheme applied to the Linear Model written in the following form:

$$c_t + V(c-b)_x = 0 \tag{3.108}$$

$$b_t = \lambda c - (\lambda+\mu)b, \tag{3.109}$$

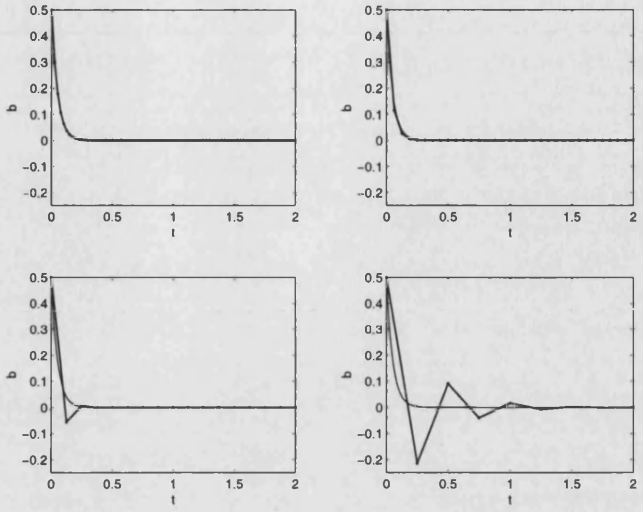

Figure 3-12: *Plot of the trapezoidal scheme applied to the first order ODE (given by dots joined by a straight line). The thin unbroken line denotes the exact solution. In all cases* $\mu = 20$. *In the top left plot* $\mu\triangle t = 0.5$, *in the top right plot* $\mu\triangle t = 1.25$, *in the bottom left plot* $\mu\triangle t = 2.5$ *and in the bottom right plot* $\mu\triangle t = 5$.

with c and b non-negative, $V > 0$ and $\lambda \geq \mu > 0$. Then the box-trap scheme applied to (3.108) and (3.109) is given by

$$\mu_x \delta_t C + p\mu_t \delta_x (C - B) = 0 \tag{3.110}$$

$$\delta_t B = \mu_t \left[\lambda' C - (\lambda' + \mu')B\right], \tag{3.111}$$

where p is the CFL number and we have again dropped the $(\cdot)_{j+\frac{1}{2}}^{n+\frac{1}{2}}$ notation for convenience. We again replace the numerical solutions C and B by their prolongations $\tilde{C}$ and $\tilde{B}$ and, as in Section 3.5.2, drop the $\tilde{(\cdot)}$ notation. Now, (3.111) can be rearranged to give B in terms of C

$$\left[\delta_t + (\lambda' + \mu')\mu_t\right] B = \lambda' \mu_t C. \tag{3.112}$$

If (3.110) is multiplied by the term in square brackets from (3.112) then we have

$$\left[\delta_t + (\lambda' + \mu')\mu_t\right] \mu_x \delta_t C + p \left[\delta_t + (\lambda' + \mu')\mu_t\right] \mu_t \delta_x (C - B) = 0. \tag{3.113}$$

Since

$$\begin{aligned} \left[\delta_t + (\lambda' + \mu')\mu_t\right] \mu_x \delta_t &= \mu_x \delta_t \left[\delta_t + (\lambda' + \mu')\mu_t\right] \\ \left[\delta_t + (\lambda' + \mu')\mu_t\right] \mu_t \delta_x &= \mu_t \delta_x \left[\delta_t + (\lambda' + \mu')\mu_t\right], \end{aligned}$$

holds, and using (3.112), the expression in (3.113) becomes

$$\mu_x\delta_t\left[\delta_t+(\lambda'+\mu')\mu_t\right]C+p\mu_t\delta_x\left[\delta_t+(\lambda'+\mu')\mu_t\right]\mu_t\delta_x C-p\mu_t\delta_x\left(\lambda'\mu_t C\right)=0. \quad (3.114)$$

Rearranging (3.114) gives an equation for C which is a quadratic in δ_t

$$\left\{\mu_x\delta_t^2+\left[(\lambda'+\mu')\mu_x\mu_t+p\mu_t\delta_x\right]\delta_t+p\mu'\mu_t^2\delta_x\right\}C=0. \quad (3.115)$$

Or, rewriting this in terms of $\mathcal{D}_t$ and $\mathcal{D}_x$ (defined in (3.90)) leads to the following quadratic equation for $\mathcal{D}_t$:

$$\mathcal{D}_t^2+(\lambda+\mu+V\mathcal{D}_x)\mathcal{D}_t+\mu V\mathcal{D}_x=0. \quad (3.116)$$

When solved, the roots must be applied to C. Note that if we had applied the box scheme to the Linear Model in terms of a and b we would have ended up with the same quadratic equation (3.116). The reason we write the Linear Model in terms of c and b in this Section is because we will compare the resulting expansion with the Improved-equilibrium model from Chapter 2 (which is derived in terms of the total concentration c).

Lemma 8. *Consider the quadratic equation (3.116) for the operator $\mathcal{D}_t$. This has solution*

$$\mathcal{D}_t=\tfrac{1}{2}\left[-(\lambda+\mu+V\mathcal{D}_x)\pm\sqrt{(\lambda+\mu+V\mathcal{D}_x)^2-4\mu V\mathcal{D}_x}\right]. \quad (3.117)$$

If the term involving the square root is expanded in increasing powers of $\mathcal{D}_x$ then the positive root gives the following expression for $\mathcal{D}_t$ in terms of $\mathcal{D}_x$ (up to and including the $\mathcal{D}_x^3$ term):

$$\mathcal{D}_t=-\frac{\mu V}{\lambda+\mu}\mathcal{D}_x+\frac{\mu\lambda V^2}{(\lambda+\mu)^3}\mathcal{D}_x^2-\frac{\mu\lambda(\lambda-\mu)V^3}{(\lambda+\mu)^5}\mathcal{D}_x^3+\ldots. \quad (3.118)$$

This gives an expression for the advected wave. Solving for the negative root leads to

$$\mathcal{D}_t=-(\lambda+\mu)-\frac{\lambda V}{\lambda+\mu}\mathcal{D}_x-\frac{\mu\lambda V^2}{(\lambda+\mu)^3}\mathcal{D}_x^2+\frac{\mu\lambda(\lambda-\mu)V^3}{(\lambda+\mu)^5}\mathcal{D}_x^3-\ldots, \quad (3.119)$$

which is an expression for the wave travelling up the time axis. The modified equation expansions are found by applying C to (3.118) and (3.119). Then the difference operators are expanded in terms of differential operators and any derivatives with respect to t are replaced by derivatives with respect to x. We then obtain a pair of modified

equation expansions; for the advected wave this is given by

$$C_t = -V'C_x + \frac{\mu\lambda V^2}{(\lambda+\mu)^3}C_{xx} - V'\left(\frac{\lambda(\lambda-\mu)V^2}{(\lambda+\mu)^4} + \frac{1}{12}[V'^2\Delta t^2 - \Delta x^2]\right)C_{xxx} + \ldots, \tag{3.120}$$

and for the wave travelling up the time axis

$$\begin{aligned} C_t &= -(\lambda+\mu)\left(1+\frac{1}{12}(\lambda+\mu)^2\Delta t^2\right)C - \bar{V}\left(1+\frac{1}{4}(\lambda+\mu)^2\Delta t^2\right)C_x \\ &\quad -\bar{V}\left(\frac{\mu V}{(\lambda+\mu)^2} + \frac{1}{12}(3\mu+2\lambda)V\Delta t^2\right)C_{xx} \\ &\quad +\bar{V}\left(\frac{\mu(\lambda-\mu)V^2}{(\lambda+\mu)^4} - \frac{1}{12}\frac{(\mu^2+3\mu\lambda+\lambda^2)V^2}{(\lambda+\mu)^2}\Delta t^2 + \frac{1}{12}\Delta x^2\right)C_{xxx} - \ldots, \end{aligned} \tag{3.121}$$

where V' is the reduced speed and $\bar{V} = V\lambda/(\lambda+\mu)$.

Proof. Let us first consider the roots in the form (3.118) and (3.119). Using the expression for $\mathcal{D}_x$ in (3.93) we can find $\mathcal{D}_x^2$ and $\mathcal{D}_x^3$ up to and including Δx^2

$$\begin{aligned} \mathcal{D}_x^2 &= \left(1 - \tfrac{1}{6}\Delta x^2\partial_x^2 + \ldots\right)\partial_x{}^2 \\ \mathcal{D}_x^3 &= \left(1 - \tfrac{1}{4}\Delta x^2\partial_x^2 + \ldots\right)\partial_x{}^3. \end{aligned}$$

We now apply to C the expressions for the differential operators $\mathcal{D}_t$ and $\mathcal{D}_x$ given in (3.118) and (3.119) and take the inverse of the bracketed term preceding ∂_t in (3.92). Then we obtain the following modified equations for the box-trap scheme applied to the linear model (up to and including the C_{xxx} term). The first is the advected wave

$$\begin{aligned} C_t &= -\frac{\mu V}{\lambda+\mu}C_x + \frac{\mu\lambda V^2}{(\lambda+\mu)^3}C_{xx} \\ &\quad -\frac{\mu V}{\lambda+\mu}\left(\frac{\lambda(\lambda-\mu)V^2}{(\lambda+\mu)^4}C_{xxx} + \frac{1}{12}\Delta t^2 C_{xtt} - \frac{1}{12}\Delta x^2 C_{xxx}\right) + \ldots, \end{aligned} \tag{3.122}$$

and the second is the wave travelling up the time axis

$$\begin{aligned} C_t &= -(\lambda+\mu)C - \frac{1}{12}(\lambda+\mu)\Delta t^2 C_{tt} - \frac{\lambda V}{\lambda+\mu}C_x - \frac{\mu\lambda V^2}{(\lambda+\mu)^3}C_{xx} \\ &\quad +\frac{\lambda V}{\lambda+\mu}\left(\frac{\mu(\lambda-\mu)V^2}{(\lambda+\mu)^4}C_{xxx} - \frac{1}{12}\Delta t^2 C_{xtt} + \frac{1}{12}\Delta x^2 C_{xxx}\right) - \ldots. \end{aligned} \tag{3.123}$$

In exactly the same way as for the linear advection equation analysed in Section 3.5.1, we wish to replace the C_{tt} and C_{xtt} terms in both (3.122) and (3.123) with C_{xx} and C_{xtt} terms respectively. The details are not given here but a simple manipulation leads

to the results (3.118) and (3.119). □

3.6.1 Discussion

The first two terms on the right hand side of (3.120) give precisely the Improved-equilibrium model which we derived in Section 2.5. In fact, the first term in the coefficient of the C_{xxx} term matches the dispersion term derived in the correction to the Improved-equilibrium model (see (A.13) in Appendix A). The derivation of the modified equation expansion is actually valid for all λ and μ whereas we had to assume that these parameters were large for the Improved-equilibrium model.

On studying (3.120) we see that there is no dependence on Δx and Δt in the diffusion term. The first coefficient that depends on the mesh is that for the dispersive term. Hence we would expect the dispersive errors to dominate and so it is the coefficient of this term that will give us information about the direction of propagation of the oscillations (which we will discuss in the next Section). This was shown to be true for the modified equation expansion of the linear advection equation (see (3.96)) where the dispersion term was dominant and the sign changed as p went through unity.

Let us neglect the Δx and Δt term (which is equivalent to taking $p' = 1$) in (3.120). Then we have (truncating after the third term)

$$C_t + \frac{\mu V}{\lambda + \mu} C_x = \frac{\mu \lambda V^2}{(\lambda + \mu)^3} C_{xx} - \frac{\lambda \mu (\lambda - \mu) V^3}{(\lambda + \mu)^5} C_{xxx}. \tag{3.124}$$

If λ and μ are equal then the dispersive term is zero. This means, as λ and μ become large, we would expect the solution to tend to the equilibrium solution (the right hand side of (3.124) would then be negligible). Now suppose $\lambda \gg \mu$; then, when both are reasonably small, there will be a large amount of diffusion (since the coefficient multiplying the C_{xx} term will be dominant). However, as they increase (but still with $\lambda \gg \mu$), there will be less reduction in the height. The dispersion coefficient will not be zero and will thus cause the pulse to be smoothed.

Figure 3-13 shows four cases of λ and μ and we are able to observe the phenomena described above. When $\lambda = \mu = 1$ (in the top left plot) the equation (3.124) does not describe the situation: the speed in the plot is V whereas (3.124) would predict the pulse to be moving at speed $V' = \frac{1}{2}$. This is simply due to the fact that, for λ and μ small, we cannot neglect the modified equation expansion up the time axis. The first term on the right hand side of (3.121) is $-(\lambda + \mu)C$. This will very quickly decay exponentially to zero as time progresses when λ and μ are large. However, it still needs to be taken into account when they are small. In the bottom right plot in Figure 3-13 we have taken $\lambda = \mu = 100$. The pulse is now moving at the reduced speed and the

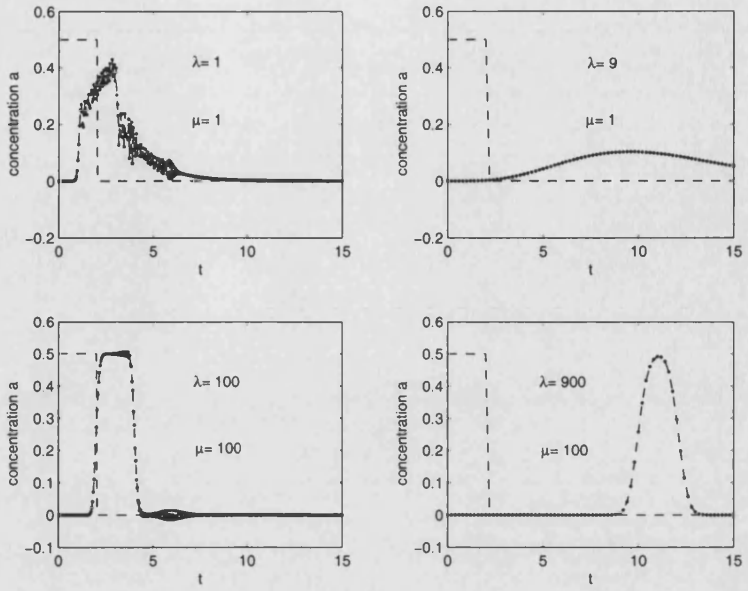

Figure 3-13: *Plot of the box-trap scheme against t (shown as dots joined by an unbroken line) applied to the linear model with* $\lambda = 90$ *and* $\mu = 10$ *at a fixed* $x = 1$ *and* $V = 1$. *The dashed line denotes the boundary condition. The values of* λ *and* μ *are:* $\lambda = \mu = 1$ *(top left),* $\lambda = \mu = 100$ *(bottom left),* $\lambda = 9$, $\mu = 1$ *(top right) and* $\lambda = 900$, $\mu = 100$ *(bottom right).*

shape of the boundary condition is preserved (to be expected since the equation is non-dispersive and the diffusion coefficient negligible). In the top left plot we observe a significant amount of diffusion when $\lambda = 9$ and $\mu = 1$. This becomes less and less as they increase to $\lambda = 900$ and $\mu = 100$ but the shape of the pulse has changed from the boundary condition.

In Figure (3-14) we have plotted the box-trap scheme applied to the Linear Model when $\lambda = \mu = 1$ for three different values of p. In the top plot the value of p is greater than one and the high frequency oscillations, which are a feature of the box scheme, propagate ahead of the wave. In the middle plot the oscillations do not spread out at all (because $p = 1$) and in the bottom plot the oscillations propagate more slowly than the main wave. This is consistent with the modified equation analysis for the linear advection equation in Section 3.5.1. It has a dispersion term which depends on the value of p. The oscillations change from propagating ahead to behind the main wave as p travels through unity because the sign of the dispersion term changes.

Hence for small λ and μ we need to use the modified equation expansion equation for the linear advection equation to describe how the oscillations behave and to choose our value of p (which we wish to take to be close to 1 to reduce the oscillations). However, for large λ and μ the modified equation expansion (3.121) describes the how the pulse

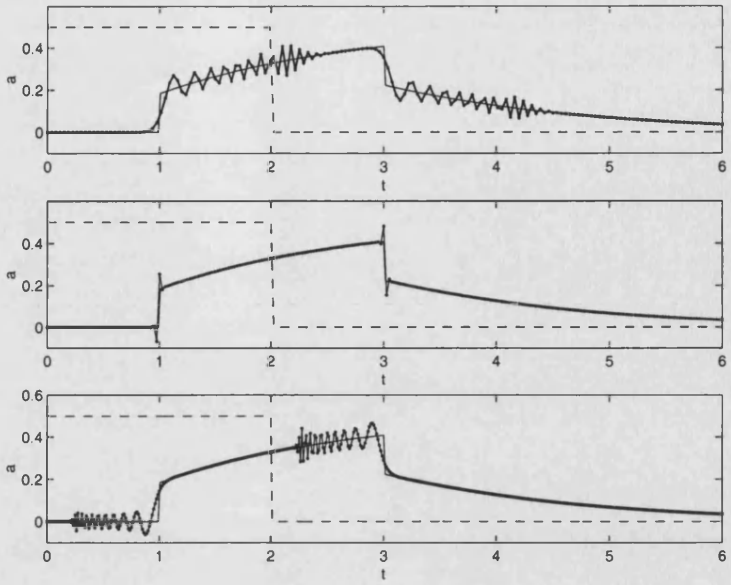

Figure 3-14: *The dashed line indicates the boundary condition, the dots joined by an unbroken line indicate the box-trap scheme and the thin unbroken line indicates the exact solution at a fixed* $x = 1$ *with* $\triangle x = 0.04$ *and* $\lambda = \mu = 1$. *In the top plot* $p = 1.5$, *in the middle plot* $p = 1$ *and in the bottom plot* $p = 0.5$.

moves and suggests we take p' close to 1 instead of p. This is not a problem when we have only one value of λ and μ in the model as we can choose either $p = 1$ or $p' = 1$ depending on their size. For larger systems it could be unclear which value to take since there could be different orders of magnitude for the parameters in the reaction terms. In Chapter 5 we will consider a larger system more closely to investigate whether the box scheme is robust enough to cope with these varying speeds.

3.6.2 Numerical experiments

In this Section we consider (3.120) without neglecting the $\triangle x$ and $\triangle t$ term in the dispersion coefficient. This term describes the variation of the advection speed of the modes with their frequencies. When λ and μ are large the propagation speed is now the reduced speed. We wish to choose $p' = 1$ to ensure the solution moves along the numerical characteristic $V'\triangle t = \triangle x$. We will refer to this as the *main wave*. As the dispersion term changes sign we should see the higher frequency oscillations change their speed relative to the speed of the main wave. Suppose we denote the coefficient of this dispersive term by Γ, i.e.

$$\Gamma = -\frac{\mu\lambda(\lambda-\mu)V^3}{(\lambda+\mu)^5} - \frac{V\mu\triangle x^2}{12(\lambda+\mu)}\left[(p')^2 - 1\right], \tag{3.125}$$

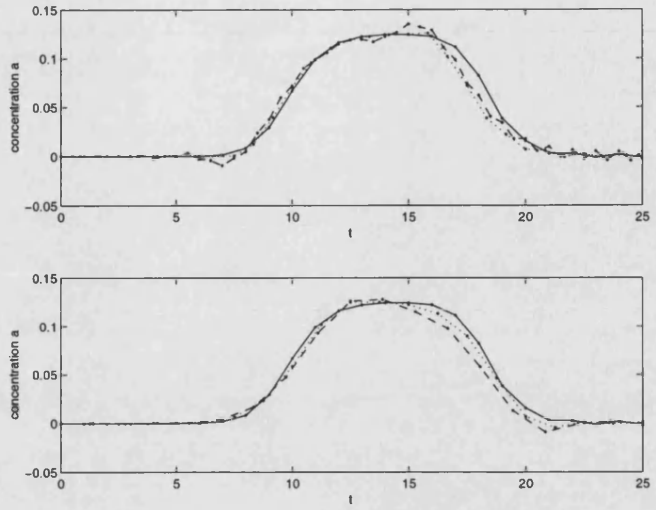

Figure 3-15: *Plot of the box-trap scheme against t applied to the linear model with* $\lambda = 90$ *and* $\mu = 10$ *at a fixed* $x = 1$. *In both plots the dots joined by an unbroken line denotes* $p' = 1$. *In the top plot* $p' = 0.8333$ *(dotted line) and* $p' = 0.5$ *(dashed line); in the bottom plot* $p' = 1.25$ *(dotted line) and* $p' = 1.3889$ *(dashed line).*

where

$$p' = \frac{V\mu}{\lambda + \mu} \frac{\triangle t}{\triangle x} \quad \left(= V' \frac{\triangle t}{\triangle x} \right). \tag{3.126}$$

In general, for $p' \geq 1$ (and $\lambda \geq \mu$) this term will be negative; but for small p' we might force a sign change. Suppose we consider the situation where $\lambda \gg \mu \gg 1$. Figure 3-15 shows the solution of the box-trap scheme with $\lambda = 90$, $\mu = 10$ for various values of p' (the boundary condition is again a square pulse). The actual value of p' that sets Γ to zero is $p' = 0.955842$. Hence for p' above this value Γ is negative and for p' below this value Γ is positive. Both plots in Figure 3-15 confirm our predictions: the change of sign in the dispersive term does give a change of direction for the oscillations. On examination of the steep sides of the wave we observe a shift as p' goes through unity: when $p' < 1$ the wave is faster than the main wave and when $p' > 1$ the wave is slower (since we are plotting against t this appears reversed in Figure 3-15).

From these plots we can observe how the smooth numerical solution and the spurious oscillatory numerical solution are separated out. The slower wave corresponds to small $k\triangle x$ and the faster wave to small $k\triangle x - \pi$. It is the latter one which is spurious and we can see that it gives oscillations both ahead and behind the wave. This will be analysed in more detail in the next Section.

In Figure 3-16 we choose a smaller $\triangle x$ and consider four different values of p'. When $p' << 1$ (top left plot) we see that there are severe oscillations. As p' increases (corresponding to an increase in $\triangle t$) these oscillations disappear but the solution becomes

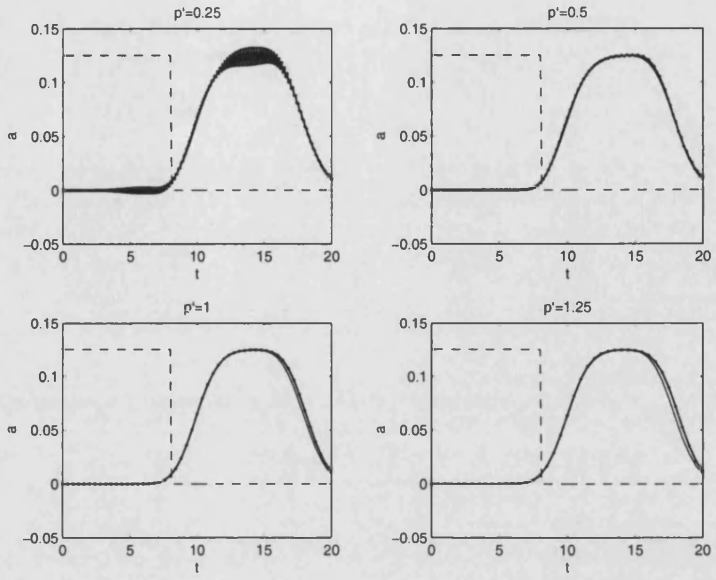

Figure 3-16: *Plot of the box-trap scheme applied to the linear model (dots joined by an unbroken line) compared with the exact Laplace transform solution (thin unbroken line). The boundary condition is a square pulse (dashed line). In all cases* $\lambda = 90$, $\mu = 10$, $\triangle x = 0.04$ *and* $x = 1$. *The values of* p' *are 0.25 (top left), 0.5 (top right), 1 (bottom left) and 1.25 (bottom right).*

more damped and we lose accuracy. The solution seems more accurate when $p' = 0.5$ rather than $p' = 1$. This is due to the fact that Γ is set to zero when $p' \approx 0.678$ for these values of the parameters. We can also observe this in Figure 3-15: the slower pulse has less oscillation but is less accurate ahead of the wave. To understand how these oscillations behave (and why they disappear as $\triangle t$ increases) we need to separately consider the main pulse and the oscillatory spurious mode.

3.6.3 Separating the smooth and oscillatory numerical solution

Consider the Linear Model given in (3.108) and (3.109). If both equations are approximated using the box scheme then there is a spurious solution mode. This is generally a problem when you use a second order difference scheme (e.g. the box scheme) to approximate a first order system. It arises from the fact that the scheme involves three time levels (it is only two time levels on each equation but when these are combined it results in a three level scheme on a single unknown) and therefore a Fourier Analysis (see Appendix B) leads to two values of the amplification factor. One of these will correspond to a good approximation of the solution of the Linear Model, but the other will not.

However, if we use the box scheme on the conservation law and the trapezoidal scheme

on the reaction equation then the oscillatory part of the solution is not totally spurious. For the full box scheme (i.e. when the box scheme is also applied to the reaction equation) the spurious solution is a solution of both equations and persists forever with no regard to the actual equation parameters. For the box-trap scheme the spurious solution gets "mixed in". Through the λ and μ terms the exact solution is smoothed out (c.f. the diffusion term in (3.120)) and the oscillatory terms can be separated out. The parameters λ and μ only have to be $O(1)$ for this smoothing to occur.

We can think of the numerical solution being made up of two parts (the smooth part and the oscillatory part) which will interact less and less as λ and μ increase. For the linear advection equation the solution is not smoothed and so we cannot separate out these two parts. We can see this from the modified equation expansion (given in (3.96)) since there is no diffusion term. This is classic "multiple time scale" analysis (Holmes 1995, pages 105-153). These problems have an oscillatory component of the solution which occurs on a time scale that is $O(1)$, and, also have a slow variation in the solution that takes place on a much smaller time scale. The two time scales are incorporated into the problem and then a power series expansion is defined in terms of these new variables. This is often known as a two-timing perturbation technique.

We now use Fourier analysis to justify the separate treatments of the smooth approximate solution and the spurious oscillatory approximate solution. Then we will be able to obtain modified equation approximations to the discrete system for each. In fact, for a purely linear problem with periodic boundary conditions, the two solutions do not interact at all. However, with our boundary conditions the modes get mixed up; though if we assume the flow is well developed they are well separated.

Let us consider the Fourier mode $A_j = e^{ikj\triangle x}$. In classical Fourier analysis we know that $\cos\frac{1}{2}k\triangle x$ is the Fourier transform (or symbol) of μ_x and, similarly, $i\sin\frac{1}{2}k\triangle x$ is the Fourier transform of δ_x. Hence

$$\mu_x A_{j+\frac{1}{2}} = \cos\tfrac{1}{2}k\triangle x A_{j+\frac{1}{2}} \tag{3.127}$$

$$\delta_x A_{j+\frac{1}{2}} = 2i\sin\tfrac{1}{2}k\triangle x A_{j+\frac{1}{2}}. \tag{3.128}$$

We wish to consider what happens when $k\triangle x$ is close to π. This is where we know that the modified equation analysis breaks down since, when we truncate, the higher order terms do not converge for $k\triangle x = \pi$. Suppose we set $k\triangle x = \pi + k'\triangle x$ where $k'\triangle x$ is small. Then (3.127) and (3.128) become

$$\mu_x A_{j+\frac{1}{2}} = -\sin\tfrac{1}{2}k'\triangle x A_{j+\frac{1}{2}} \tag{3.129}$$

$$\delta_x A_{j+\frac{1}{2}} = 2i\cos\tfrac{1}{2}k'\triangle x A_{j+\frac{1}{2}}. \tag{3.130}$$

Write

$$\begin{aligned} A_j = e^{ij(\pi+k'\Delta x)} &= (-1)^j e^{ik'j\Delta x} \\ &\equiv (-1)^j A_j^o, \end{aligned} \tag{3.131}$$

where A^o denotes the oscillatory part of the solution. Substituting this into (3.129) gives

$$\begin{aligned} \mu_x A_{j+\frac{1}{2}} &= -\sin\tfrac{1}{2}k'\Delta x(-1)^{j+\frac{1}{2}} A^o_{j+\frac{1}{2}} \\ &= (-1)^{j+1}\tfrac{1}{2}\delta_x A^o_{j+\frac{1}{2}}, \end{aligned} \tag{3.132}$$

and similarly

$$\delta_x A_{j+\frac{1}{2}} = (-1)^{j+1} 2\mu_x A^o_{j+\frac{1}{2}}. \tag{3.133}$$

Now suppose $A_j^n = (-1)^{j+n}(A^o)_j^n$, which is the spurious oscillatory approximate solution. Then

$$\mu_x A^n_{j+\frac{1}{2}} = (-1)^{j+n+1}\tfrac{1}{2}\delta_x (A^o)^n_{j+\frac{1}{2}}, \tag{3.134}$$

$$\delta_x A^n_{j+\frac{1}{2}} = (-1)^{j+n+1} 2\mu_x (A^o)^n_{j+\frac{1}{2}}, \tag{3.135}$$

which are analogous to (3.132) and (3.133) respectively. We can apply exactly the same analysis to the difference operators μ_t and δ_t to give

$$\mu_t A_j^{n+\frac{1}{2}} = (-1)^{j+n+1}\tfrac{1}{2}\delta_t (A^o)_j^{n+\frac{1}{2}} \tag{3.136}$$

$$\delta_t A_j^{n+\frac{1}{2}} = (-1)^{j+n+1} 2\mu_t (A^o)_j^{n+\frac{1}{2}}. \tag{3.137}$$

We wish to apply this analysis to the box-trap scheme for the Linear Model in (3.40) and (3.42) (but with $A+B$ in (3.40) replaced by C). The discretisation involves the operators $\mu_x\delta_t$ and $\mu_t\delta_x$ and so, applying μ_x to (3.137) and μ_t to (3.135) gives

$$\mu_x\delta_t A^{n+\frac{1}{2}}_{j+\frac{1}{2}} = (-1)^{j+n+2}\delta_x\mu_t (A^o)^{n+\frac{1}{2}}_{j+\frac{1}{2}} \tag{3.138}$$

$$\mu_t\delta_x A^{n+\frac{1}{2}}_{j+\frac{1}{2}} = (-1)^{j+n+2}\delta_t\mu_x (A^o)^{n+\frac{1}{2}}_{j+\frac{1}{2}}. \tag{3.139}$$

Suppose the numerical solution is separated as follows:

$$A_j^n \equiv (A^s)_j^n + (-1)^{j+n}(A^o)_j^n, \tag{3.140}$$

and similarly for B_j^n and C_j^n. This means that $(\cdot)^s$ represents the smooth part of the solution and $(\cdot)^o$ represents the smooth modulation of the spurious oscillatory solution. Assuming we can consider the smooth and oscillatory parts separately, we can substitute these into (3.40) and (3.42) and then apply the expressions in (3.138)

and (3.139) to the oscillatory part. This leads to two systems for the $(\cdot)^s$ and $(\cdot)^o$ parts of the solution, namely

$$\mu_x\delta_t C^s + p\mu_t\delta_x A^s \approx 0 \tag{3.141}$$

$$\delta_t B^s \approx \mu_t\left(\lambda' A^s - \mu' B^s\right), \tag{3.142}$$

and

$$\delta_x\mu_t C^o + p\delta_t\mu_x A^o \approx 0 \tag{3.143}$$

$$-2\mu_t B^o \approx -\tfrac{1}{2}\delta_t\left(\lambda' A^o - \mu' B^o\right). \tag{3.144}$$

Multiplying both sides of (3.143) and (3.144) by $\mu/(\lambda+\mu)$ gives

$$\frac{\mu}{\lambda+\mu}\delta_x\mu_t C^o + \frac{\mu p}{\lambda+\mu}\delta_t\mu_x A^o \approx 0, \tag{3.145}$$

and

$$-\frac{2\mu}{\lambda+\mu}\mu_t B^o \approx -\tfrac{1}{2}\frac{\mu}{\lambda+\mu}\delta_t\left(\lambda' A^o - \mu' B^o\right). \tag{3.146}$$

If we define a new variable D^o by

$$D^o = \frac{\lambda A^o - \mu B^o}{\lambda+\mu}, \tag{3.147}$$

then we can eliminate C^o and B^o from (3.145) and (3.146) to obtain a system in terms of A^o and D^o

$$\delta_t\mu_x A^o + \frac{1}{p'}\delta_x\mu_t\left(A^o - D^o\right) \approx 0 \tag{3.148}$$

$$\delta_t D^o \approx \frac{4\lambda}{\mu(\lambda'+\mu')}\mu_t A^o - \frac{4}{\mu'}\mu_t D^o, \tag{3.149}$$

where $p' = \mu p/(\lambda+\mu)$. Also introduce new variables

$$\bar{\lambda} = \frac{4\lambda}{\mu(\lambda'+\mu')}, \quad \bar{\mu} = \frac{4}{\lambda'+\mu'}, \quad \bar{p} = \frac{1}{p'}. \tag{3.150}$$

Then $\bar{\lambda}+\bar{\mu} = 4/\mu'$ and so (3.148) and (3.149) become

$$\mu_x\delta_t A^o + \bar{p}\mu_t\delta_x(A^o - D^o) \approx 0 \tag{3.151}$$

$$\delta_t D^o \approx \mu_t\left(\bar{\lambda}A^o - (\bar{\mu}+\bar{\lambda})D^o\right). \tag{3.152}$$

These are of the same form as (3.110) and (3.111) but for A^o and D^o instead of C and B and with p, λ' and $\lambda'+\mu'$ replaced by $\bar{p}$, $\bar{\lambda}$ and $\bar{\lambda}+\bar{\mu}$ respectively.

In Section 3.6 we found the modified equation expansions by solving a quadratic in $\mathcal{D}_t$,

given in (3.115), with the roots applied to C. Hence we can now obtain an analogous equation which is applied to A^o

$$\mathcal{D}_t^2 + \left[\frac{\bar{\lambda}+\bar{\mu}}{\Delta t} + \frac{\bar{p}\Delta x}{\Delta t}\mathcal{D}_x\right]\mathcal{D}_t + \frac{\bar{\mu}\bar{p}\Delta x}{\Delta t^2}\mathcal{D}_x = 0. \tag{3.153}$$

Finally, define new variables

$$\Lambda := \frac{\bar{\lambda}}{\Delta t}, \quad M := \frac{\bar{\mu}}{\Delta t}, \quad W := \frac{\bar{p}\Delta x}{\Delta t}. \tag{3.154}$$

Then we have a quadratic to solve for $\mathcal{D}_t$ which is precisely of the form of (3.116)

$$\mathcal{D}_t^2 + [(\Lambda + M) + W\mathcal{D}_x]\mathcal{D}_t + MW\mathcal{D}_x = 0. \tag{3.155}$$

One of the roots of (3.155) will give a modified equation expansion for the wave travelling up the time axis and the other will give a modified equation expansion for the advected wave. The positive root (i.e. the advected wave) will give the following expression for $\mathcal{D}_t$ in terms of $\mathcal{D}_x$ (up to and including the $\mathcal{D}_x^3$ term):

$$\mathcal{D}_t = -\frac{MW}{\Lambda+M}\mathcal{D}_x + \frac{M\Lambda W^2}{(\Lambda+M)^3}\mathcal{D}_x^2 - \frac{M\Lambda(\Lambda-M)W^3}{(\Lambda+M)^5}\mathcal{D}_x^3 + \ldots. \tag{3.156}$$

Following the procedure in Section 3.6 we can substitute for $\mathcal{D}_x$, $\mathcal{D}_x^2$ and $\mathcal{D}_x^3$ etc. and apply this expression to A^o. We will eventually obtain an expression for $(A^o)_t$ entirely in terms of x-derivatives of A^o for this advected wave. The coefficients of the $(A^o)_x$ and $(A^o)_{xx}$ terms will be the same as the coefficients multiplying $\mathcal{D}_x$ and $\mathcal{D}_x^2$ respectively. A simple calculation gives

$$\frac{MW}{\Lambda+M} = \frac{V}{p^2}, \qquad \frac{M\Lambda W^2}{(\Lambda+M)^3} = \frac{\lambda}{4V^2}\Delta x^4\Delta t^2. \tag{3.157}$$

It is interesting to note that the first expression in (3.157) is independent of λ and μ. Also, the coefficient of the damping term (i.e. the second expression) increases as Δt and Δx increase. This explains why the oscillations in Figures 3-16 and 3-15 decreased as Δt increased for fixed Δx: as Δt gets larger the oscillations become more and more damped.

This analysis also gives us information about how the envelope of oscillations move. We can see how the coefficient of the $(A^o)_x$ term (3.157) will give us the position of these oscillations by looking at some numerical results. Figure 3-17 confirms these predictions. On examining (3.157), we see that the oscillations should move with speed $\frac{V}{p^2}$. In the left three plots in Figure 3-17 we have chosen $V = 1$ and observe that the oscillations first occur at time $\frac{p^2 x}{V}$, i.e. $t = 4$. The second set of oscillations are

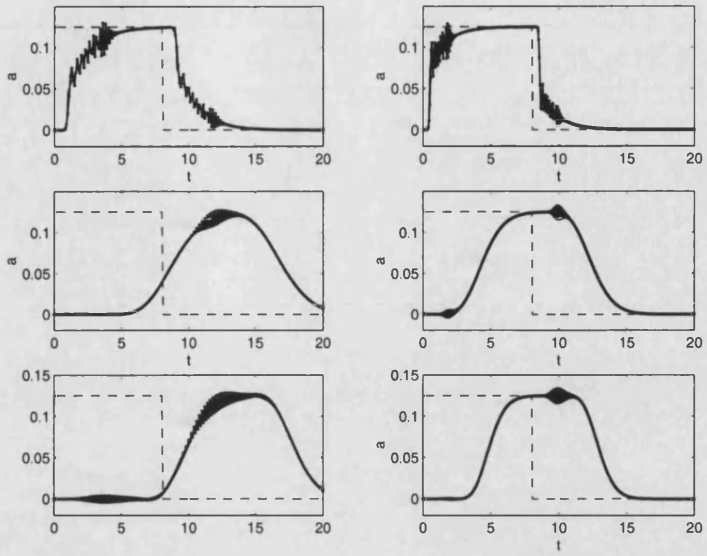

Figure 3-17: *The dashed line indicates the boundary condition, the dots joined by an unbroken line indicates the numerical solution and the thin unbroken line indicates the exact solution at a fixed distance* $x = 1$ *with* $p = 2$, $\triangle x = 0.05$ *and* $\lambda = \mu = 1$ *(top plots),* $\lambda = 40$ *and* $\mu = 5$ *(middle plots) and* $\lambda = 90$ *and* $\mu = 10$ *(bottom plots). Also, in the left three plots* $V = 1$ *and in the right three plots* $V = 2$.

expected to occur at the width of the boundary condition after the first set, which is $t = 12$. In the right three plots we have taken $V = 2$ so the the first set of oscillations are expected to occur at $t = 2$ (and therefore the second set at $t = 10$). We also see that they are independent of the choice of λ and μ since the oscillations are in the same position for all three cases.

3.6.4 Modified equation analysis of the weighted box-trap scheme

Finally in this Chapter we outline the derivation to obtain the modified equation expansions of the weighted box-trap scheme applied to the Linear Model. The discretised system is given by (3.110) and (3.111) with μ_t replaced by θ_t in (3.110). Now, θ_t can be expanded in terms of differential operators to give

$$\theta_t = 1 + \left(\theta - \tfrac{1}{2}\right) \triangle t \partial_t + \tfrac{1}{8} \triangle t^2 \partial_t^2 + \ldots. \tag{3.158}$$

A similar analysis leads to the following quadratic equation (c.f. equation (3.116)) but with $\mathcal{D}_t$ replaced by $\tilde{\mathcal{D}}_t$, and an extra operator, $\mathcal{M}_t$, i.e.

$$\tilde{\mathcal{D}}_t^2 + \left[(\lambda + \mu)\mathcal{M}_t + V\mathcal{D}_x\right]\tilde{\mathcal{D}}_t + \mu V \mathcal{M}_t \mathcal{D}_x = 0, \tag{3.159}$$

	l^2 norm	maximum error
$p' = 0.25,\ \theta = 0.5$	0.047066	0.007616
$p' = 0.25,\ \theta = 0.51$	0.016018	0.003236
$p' = 0.1,\ \theta = 0.5$	0.018354	0.007640
$p' = 0.1,\ \theta = 0.51$	0.018025	0.006747
$p' = 0.64103,\ \theta = 0.51$	0.008734	0.002249

Table 3.1: *Table showing a comparison of the l^2 norm and the maximum errors between the exact solution and the weighted box-trap scheme for the examples in Figure 3-18 with one extra case, the fourth row, not plotted.*

where

$$\tilde{\mathcal{D}}_t := \frac{\theta_t^{-1}\delta_t}{\triangle t}, \qquad \mathcal{M}_t := \theta_t^{-1}\mu_t. \tag{3.160}$$

After some analysis, the modified equation expansion is found to be (considering the expansion for the advected wave only)

$$\begin{aligned} C_t \;=\; & -\frac{\mu V}{\lambda+\mu}C_x + \frac{\mu V}{\lambda+\mu}\left\{\frac{\lambda V}{(\lambda+\mu)^2} + \frac{\mu V}{\lambda+\mu}\left(\theta-\tfrac{1}{2}\right)\triangle t\right\}C_{xx} \\ & -\frac{\mu V}{\lambda+\mu}\left\{\frac{\lambda(\lambda-\mu)V^2}{(\lambda+\mu)^4} + \frac{1}{12}\left[\frac{\mu^2V^2}{(\lambda+\mu)^2}\triangle t^2 - \triangle x^2\right]\right. \\ & \left. + \frac{\mu^2V^2}{(\lambda+\mu)^2}(\theta-\tfrac{1}{2})^2\triangle t^2 + 3\frac{\lambda\mu V^2}{(\lambda+\mu)^3}(\theta-\tfrac{1}{2})\triangle t\right\}C_{xxx} + \ldots. \end{aligned} \tag{3.161}$$

On comparing the expansion (3.161) with (3.123) we see that there is now dependence on $\triangle t$ in the diffusion term. This is to be expected since numerical simulations have already shown that weighted box-trap scheme introduces extra diffusion. However, the extra term $\frac{\mu V}{\lambda+\mu}\left(\theta-\frac{1}{2}\right)\triangle t$ will be small if $\theta = \frac{1}{2} + O(\triangle t)$.

In the Discussion in Section 3.6.1 we stated that, for large λ and μ, we wish to take $p' = 1$, as guided by the modified equation expansion corresponding to the advected wave (the wave travelling up the time axis can be ignored in this case). However, the results in Figure 3-16 illustrated that the numerical solution was most accurate when p' was chosen so that $\Gamma = 0$ (defined in 3.125). If λ and μ are sufficiently large this is effectively $p' = 1$ but we will now expand on the remarks in Section 3.6.2 to show that there are also situations where this is not the case for the weighted box-trap scheme. Suppose $\triangle x$ and θ are fixed and then the dispersion term in (3.161) is set to zero. This gives the following quadratic equation for $\triangle t$ which is solved (choosing the positive root):

$$\frac{\mu^2V^2}{(\lambda+\mu)^2}\left[\epsilon^2 + \frac{1}{12}\right]\triangle t^2 + 3\frac{\lambda\mu V^2}{(\lambda+\mu)^3}\epsilon\triangle t + \frac{\lambda(\lambda-\mu)V^2}{(\lambda+\mu)^4} - \frac{1}{12}\triangle x^2 = 0, \tag{3.162}$$

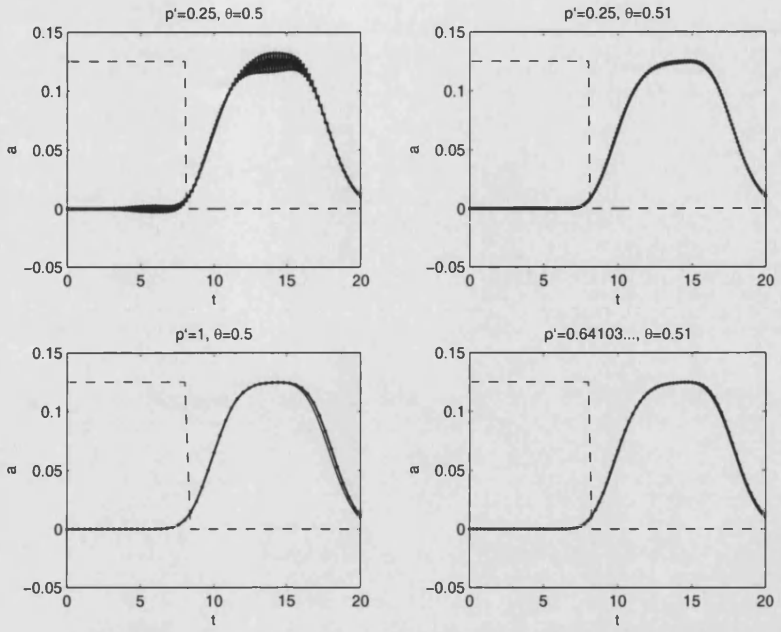

Figure 3-18: *The broken line indicates the boundary condition, the dots joined by an unbroken line indicate the numerical solution and the thin unbroken line indicates the exact Laplace transform solution with* $\lambda = 90$, $\mu = 10$ *and* $\triangle x = 0.04$. *Other parameter values are:* $p' = 0.25$ *and* $\theta = 0.5$ *(top left),* $p' = 0.25$ *and* $\theta = 0.51$ *(top right),* $p' = 1$ *and* $\theta = 0.5$ *(bottom left),* $p' = 0.64103...$ *and* $\theta = 0.51$ *(bottom right).*

where $\epsilon = \theta - \frac{1}{2}$. In Figure 3-18 we have plotted the weighted box-trap scheme with $\lambda = 90$ and $\mu = 10$ for various p' $(= V'\triangle t/\triangle x)$. In the top two plots $p' = 0.25$ and we observe that the oscillations are severe (top left). These are reduced by increasing θ from 0.5 to 0.51 but we would like to be able to take larger time-steps. When $p' = 1$ and $\theta = 0.5$ (in the bottom left plot) the oscillations have disappeared but accuracy has been lost. Although not shown here, if θ is increased to 0.51 the results are very similar. However, in the bottom right plot we have chosen $p' = 0.64103$ which sets the dispersion term to zero. The results here are much more accurate and so gives a very effective way to choose $\triangle t$ to eliminate the oscillations and maintain accuracy. This is confirmed in Table 3.1 where the l^2 norm and maximum error have been found for the parameter values in Figure 3-18. The entries in the final row, corresponding to the bottom right plot in Figure 3-18, are significantly smaller.

Chapter 4

The box scheme for nonlinear conservation laws

4.1 Introduction

In this Chapter we apply the box scheme to nonlinear conservation laws. These arise in problems of interest here when considering a two equation model with a nonlinear reaction term. In the Introduction (Section 1.3.2) we described the Langmuir Model which falls into this category. Suppose we write the equations in the following form:

$$a_t + b_t + V a_x = 0 \tag{4.1}$$

$$b_t = \lambda a(B - b) - \mu b. \tag{4.2}$$

In equilibrium we assume that the right hand side of (4.2) is zero and so

$$b = g(a) = \frac{BKa}{1 + Ka}, \tag{4.3}$$

where $K := \lambda/\mu$. This is called the *Langmuir Isotherm.* Then (4.1) becomes

$$a_t + \left(\frac{V}{1 + g'(a)}\right) a_x = 0. \tag{4.4}$$

We have already seen that the box scheme applied to linear conservation laws gives oscillations when the boundary data is non-smooth which can be reduced by weighting in the t direction. We now wish to investigate how the box scheme copes with nonlinear flux functions when the initial data is a shock, or when a shock forms.

Consider a general nonlinear conservation law in conservative form

$$u_t + f(u)_x = 0. \tag{4.5}$$

Our aim is to try to understand the box scheme applied to this problem in the presence of shocks. In this Chapter we will use Burgers' equation, where $f(u) := \frac{1}{2}u^2$, as our model problem. This should be called the *inviscid Burgers' equation* as it involves no diffusion term but, for convenience, we will refer to it as Burgers' equation from now on. Suppose we have Riemann data

$$u(x,0) = \begin{cases} u_l, & x < \sigma \\ u_r, & x > \sigma, \end{cases} \tag{4.6}$$

for all $0 \leq x \leq X$ with $u_l > u_r$. Also assume that $u_r \geq 0$ so that boundary data only needs to be specified on the left boundary, i.e.

$$u(0,t) = u_l, \tag{4.7}$$

for all $0 \leq t \leq T$. The exact solution is given by

$$u(x,t) = \begin{cases} u_l, & x < \sigma + st \\ u_r, & x > \sigma + st, \end{cases} \tag{4.8}$$

where s is the shock speed

$$s = \tfrac{1}{2}(u_l + u_r). \tag{4.9}$$

The box scheme applied to (4.5) in conservative form is defined to be

$$\delta_t \mu_x U_{j+\frac{1}{2}}^{n+\frac{1}{2}} + \nu \delta_x \mu_t F_{j+\frac{1}{2}}^{n+\frac{1}{2}} = 0, \tag{4.10}$$

where $F_j^n := f\left(U_j^n\right)$ and $\nu := \Delta t / \Delta x$ is the mesh ratio. Figure 4-1 show plots of the box scheme for Burgers' equation after one time-step (on the left) and ten time-steps (on the right) for two values of u_l and u_r. We can see that the box scheme gives oscillations even after one time-step and these increase in both size and number as time progresses. We would like to modify the box scheme to eliminate these oscillations. Our aims in this chapter are

1. to investigate how well the box scheme approximates the exact solution of non-linear conservation laws with non-smooth data,

2. to understand how these oscillations arise when applying the box scheme to a nonlinear conservation law in the presence of a shock, and

3. to explore post-processing the results to eliminate the oscillations before moving to the next time-step.

The box scheme can be interpreted as having a final projection stage: since the underlying approximation can be taken as piecewise linear or bilinear and the "test functions"

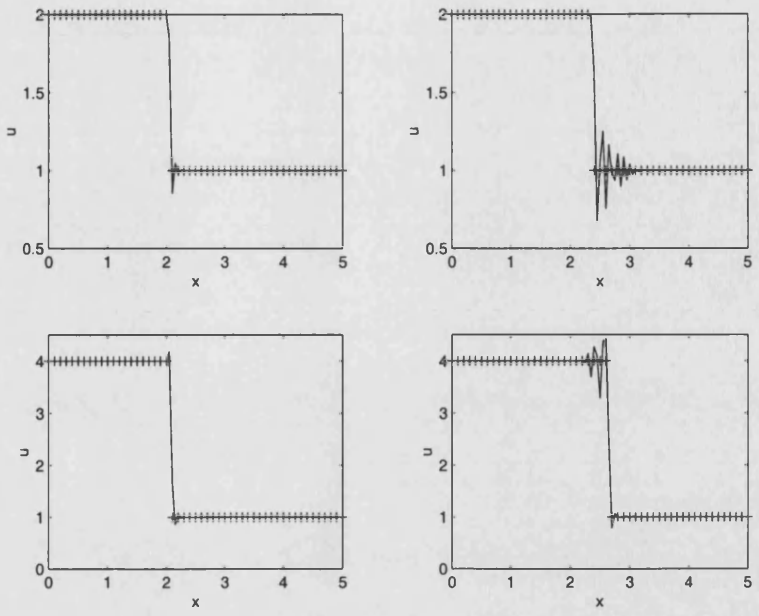

Figure 4-1: *Plot of the box scheme for Burgers' equation (straight line) compared with the exact solution (+'s) with* $\nu := \triangle t/\triangle x = 0.5$. *In the top two plots* $u_l = 2$, $u_r = 1$ *after one time-step (left) and ten time-steps (right) and in the bottom two plots* $u_l = 4$, $u_r = 1$ *after one time-step (left) and ten time-steps (right).*

are piecewise constant, it has some of the characteristics of a Petrov-Galerkin (P-G) method. As described in detail in (Morton 1996), in P-G methods the trial space and test space are not necessarily the same, whereas in Galerkin methods the same space of functions are used for both the approximate solutions and the weighting functions. From (4.8) we know that the theoretical solution of the Riemann problem for Burgers' equation is a shock which propagates to the right as time progresses. We will firstly investigate how the P-G method deals with discontinuities and then derive the box scheme using the trapezoidal rule to approximate the integrals that arise in this P-G approach.

4.2 Approximation of a given function using a P-G method

In this section we wish to approximate a given function $u(x)$ using the P-G method which has a piecewise linear trial space and a piecewise constant test space. Consider a grid $0 = x_0 < x_1 < \ldots < x_J = 1$ and denote the subinterval I_j by $I_j = (x_{j-1}, x_j)$ for $j = 1, \ldots, J$. We use a constant step length, h, i.e. $h = x_j - x_{j-1}$ for all $j = 1, \ldots, J$.

Define the piecewise linear approximation U to a function u as

$$U(x) = \sum_{j=0}^{J} U_j \phi_j(x), \tag{4.11}$$

where we seek to determine the coefficients U_j. Let $\phi_j(x)$, $j = 0, \ldots, J$, be piecewise linear basis functions given by

$$\phi_j(x) = \begin{cases} \frac{x - x_j}{h}, & x \in I_j := (x_{j-1}, x_j) \\ \frac{x_{j+1} - x}{h}, & x \in I_{j+1} \cup \{x_j\} := [x_j, x_{j+1}), \\ 0, & \text{otherwise,} \end{cases} \tag{4.12}$$

for $j = 1, \ldots, J-1$, with

$$\phi_0(x) = \frac{x_1 - x}{h}, \quad \phi_J(x) = \frac{x - x_{J-1}}{h}. \tag{4.13}$$

Then $U_j = U(x_j)$. Now define the piecewise constant test functions by

$$\chi_i(x) = \begin{cases} 1, & x \in (x_{i-1}, x_i) \\ 0, & \text{otherwise,} \end{cases} \tag{4.14}$$

for $i = 1, \ldots, J$. A P-G method for finding U_j requires that the difference $U(x) - u(x)$ be orthogonal to these test functions χ_i, i.e.

$$\langle U(x) - u(x), \chi_i(x) \rangle = 0, \quad i = 1, \ldots, J, \tag{4.15}$$

where we define $\langle \cdot, \cdot \rangle$ as the l^2 inner product

$$\langle u, v \rangle = \int_0^1 u(x) v(x) \, \mathrm{d}x. \tag{4.16}$$

Substituting (4.11) into (4.15) gives

$$\Big\langle \sum_{j=0}^{J} U_j \phi_j, \chi_i \Big\rangle = \langle u, \chi_i \rangle, \quad i = 1, \ldots, J,$$

or, by linearity of the l^2 inner product

$$\sum_{j=0}^{J} U_j \langle \phi_j, \chi_i \rangle = \langle u, \chi_i \rangle, \quad i = 1, \ldots, J. \tag{4.17}$$

The set of equations defined in (4.17) can be written as the matrix system

$$K\mathbf{U} = \mathbf{g}, \tag{4.18}$$

where $K \in \mathbb{R}^{J\times(J+1)}$, $\mathbf{U} \in \mathbb{R}^{(J+1)\times 1}$ and $\mathbf{g} \in \mathbb{R}^{J\times 1}$; so, componentwise we have

$$\begin{aligned} K_{ij} &= \langle \phi_j, \chi_i \rangle, \quad j = 0, \ldots, J, \quad \& \quad i = 1, \ldots, J, \\ g_i &= \langle u, \chi_i \rangle, \quad i = 1, \ldots, J. \end{aligned}$$

Now,

$$\langle \phi_j, \chi_i \rangle = \begin{cases} \frac{1}{2}h, & j = i \\ \frac{1}{2}h, & j = i-1 \\ 0, & i = k+1, \ldots, J, \end{cases} \tag{4.19}$$

and so

$$K = \tfrac{1}{2}h \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 \\ 0 & 1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & & \ddots & 1 & 1 \end{pmatrix}. \tag{4.20}$$

Note that (4.18) is a system of J equations and $J+1$ unknowns so we will need a boundary value to solve for U_j.

4.2.1 A simple example

Suppose u is a given function which is discontinuous at one point. A piecewise linear U is sought as an approximation by a P-G method. The nodal projection, which can be regarded as a P-G method with the test functions comprising the set of delta functions at the nodes, gives no information about the shock position within the cell; whereas the P-G method based on the test functions defined in (4.14) gives information from which the shock position can be deduced. Define u as

$$u(x) = \begin{cases} 1, & x < \sigma \\ 0, & x > \sigma, \end{cases} \tag{4.21}$$

which corresponds to a shock at $x = \sigma$. The nodal values are simply

$$u(x_i) = \begin{cases} 1, & i = 0, \ldots, k \\ 0, & i = k+1, \ldots, J, \end{cases} \tag{4.22}$$

provided $x_k < \sigma < x_{k+1}$, and we write

$$\sigma = x_k + \alpha h, \tag{4.23}$$

for some $\alpha \in (0, 1)$. So the nodal values gives the value of k but no information on α.

In the P-G method using piecewise constant test functions, the prescribed left hand

boundary value gives the coefficient $U_0 = u(x_0) = 1$ and so (4.18) becomes a system of J equations in J unknowns. Now

$$(u, \chi_i) = \begin{cases} h, & i = 1, \ldots, k \\ \alpha h, & i = k+1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus

$$\mathbf{g} = h(1, \ldots, 1, \alpha, 0, \ldots, 0)^T. \tag{4.24}$$

We can extend K in (4.18) to be square by adding the value $U_0 = 1$. Then

$$\tilde{K}\mathbf{U} = \tilde{\mathbf{g}}, \tag{4.25}$$

where $\tilde{K} \in \mathbb{R}^{(J+1)\times(J+1)}$ and $\tilde{\mathbf{g}} \in \mathbb{R}^{(J+1)\times 1}$ and we have

$$\tfrac{1}{2}\begin{pmatrix} 1 & 0 & \ldots & & 0 \\ 1 & 1 & \ddots & & \vdots \\ 0 & 1 & 1 & \ddots & \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & & & 1 & 1 \end{pmatrix}\begin{pmatrix} U_0 \\ U_1 \\ \vdots \\ \vdots \\ U_J \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 1 \\ \vdots \\ 1 \\ \alpha \\ 0 \\ \vdots \end{pmatrix}, \tag{4.26}$$

i.e. $\tilde{K}$ and $\tilde{\mathbf{g}}$ have had h scaled out. Writing $\tilde{K} = \frac{1}{2}(I + S)$, where S is the null matrix with ones on the first subdiagonal, so that

$$\tilde{K}^{-1} = 2(I + S)^{-1} = 2(I - S + S^2 - S^3 + \ldots),$$

we obtain

$$\tilde{K}^{-1} = 2\begin{pmatrix} 1 & 0 & \ldots & & 0 \\ -1 & 1 & \ddots & & \vdots \\ 1 & -1 & 1 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \\ & & 1 & -1 & 1 \end{pmatrix}. \tag{4.27}$$

Thus (4.26) can solved to give

$$U_j = \begin{cases} 1, & j = 0, \ldots, k \\ (-1)^{j-k-1}(2\alpha - 1), & j = k+1, \ldots, J. \end{cases} \tag{4.28}$$

This is oscillatory for $j = k+1, \ldots, J$. Hence the P-G approach, using piecewise linear basis functions ϕ_j and piecewise constant test functions χ_i, gives coefficients with an alternating structure after the discontinuity and the oscillation tells us where the shock is within the cell.

Suppose we now consider a discontinuous function of the form

$$u(x) = \begin{cases} u_l, & x < \sigma \\ u_r, & x > \sigma, \end{cases} \tag{4.29}$$

and so the nodal values are

$$u(x) = \begin{cases} u_l, & i = 0, \ldots, k \\ u_r, & i = k+1, \ldots, J, \end{cases} \tag{4.30}$$

If we again assume that $\sigma = x_k + \alpha h$, for some $\alpha \in (0,1)$ then

$$(u, \chi_i) = \begin{cases} u_l h, & i = 1, \ldots, k \\ \alpha u_l h + (1-\alpha) u_r h, & i = k+1 \\ u_r h, & i = k+2, \ldots, J. \end{cases}$$

Setting $U_0 = u_l$ and applying the same procedure as above gives (c.f.(4.28))

$$U_j = \begin{cases} u_l, & j = 0, \ldots, k \\ (-1)^{j-k-1}\big[(2\alpha-1)u_l + (1-2\alpha)u_r + u_r\big], & j = k+1, \ldots, J, \end{cases} \tag{4.31}$$

which is also oscillatory for $j = k+1, \ldots, J$ about u_r. We can recover u_r from $\mu_x U_{j+\frac{1}{2}}$ for $j = k+1, \ldots, J$ since

$$\mu_x U_{j+\frac{1}{2}} = \tfrac{1}{2}(U_j + U_{j+1}) = u_r, \tag{4.32}$$

for $j = k+1, \ldots, J$. Provided we know u_l (e.g. from the boundary value), we can also deduce α from $\delta_x U_{j+\frac{1}{2}}$ for $j = k+1, \ldots, J$ using

$$\begin{aligned} \delta_x U_{j+\frac{1}{2}} = U_{j+1} - U_j &= 2(2\alpha-1)u_l + 2(1-2\alpha)u_r \\ &= 2(2\alpha-1)(u_l - u_r), \end{aligned} \tag{4.33}$$

for $j = k+1, \ldots, J$. Note that these are the same oscillations as for the case when $u_l = 1$ and $u_r = 0$ but they have now been shifted and scaled.

4.3 Derivation of the box scheme as a P-G method

4.3.1 The time independent problem

Consider the following ordinary differential equation on $x \in (0,1)$:

$$u'(x) = S(u), \quad u(0) = u_0. \tag{4.34}$$

This is of the form

$$Lu = S, \tag{4.35}$$

where

$$Lu := \{u', \text{ with } u(0) = u_0\}. \tag{4.36}$$

We wish to find a piecewise linear approximation to the solution of (4.34). Hence we seek a solution

$$U(x) = \sum_{j=0}^{J} U_j \phi_j(x), \tag{4.37}$$

where $\phi_j(x)$ are linear basis functions defined in (4.12) and $U_0 = u_0$. Our P-G method requires that the residual $LU(x) - S$ is orthogonal to the piecewise constant test functions $\chi_i(x)$ defined in (4.14), and so

$$\langle LU(x) - S(U(x)), \chi_i(x) \rangle = 0, \quad i = 1, \ldots, J. \tag{4.38}$$

Now

$$LU(x) = U'(x) = \sum_{j=0}^{J} U_j \phi_j'(x), \tag{4.39}$$

and so (4.38) becomes

$$\sum_{j=0}^{J} U_j \langle \phi_j', \chi_i \rangle = \langle S, \chi_i \rangle, \quad i = 1, \ldots, J. \tag{4.40}$$

Also, since

$$\phi_j'(x) = \begin{cases} \frac{1}{h}, & x \in I_j := (x_{j-1}, x_j) \\ -\frac{1}{h}, & x \in I_{j+1} := (x_j, x_{j+1}), \\ 0, & \text{otherwise,} \end{cases} \tag{4.41}$$

for $j = 1, \ldots, J-1$, with

$$\phi_0'(x) = -\frac{1}{h}, \qquad \phi_J'(x) = \frac{1}{h}, \tag{4.42}$$

we again obtain a matrix system $K\mathbf{U} = \mathbf{g}$, where

$$\begin{aligned} K_{ij} &= \langle \phi_j', \chi_i \rangle, \quad j = 0, \ldots, J, \quad \& \quad i = 1, \ldots, J, \\ g_i &= \langle S, \chi_i \rangle, \quad i = 1, \ldots, J, \end{aligned}$$

and

$$\langle \phi_j', \chi_i \rangle = \begin{cases} 1, & j = i \\ -1, & j = i - 1 \\ 0, & \text{otherwise.} \end{cases} \tag{4.43}$$

This gives

$$K = \begin{pmatrix} -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & & \ddots & -1 & 1 \end{pmatrix}. \tag{4.44}$$

Also

$$\langle S, \chi_i \rangle = \int_{x_{i-1}}^{x_i} S(U)\, \mathrm{d}x =: I_i, \qquad i = 1, \ldots, J, \tag{4.45}$$

and we have set $U_0 = u_0$. The system to solve is now

$$\begin{pmatrix} 1 & 0 & \ldots & 0 \\ -1 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \\ \vdots & & -1 & 1 \end{pmatrix} \begin{pmatrix} U_1 \\ \vdots \\ \vdots \\ U_J \end{pmatrix} = \begin{pmatrix} u_0 + I_1 \\ I_2 \\ \vdots \\ I_J \end{pmatrix}. \tag{4.46}$$

Suppose the integral I_i is approximated by the trapezoidal rule, i.e.

$$I_i = \int_{x_{i-1}}^{x_i} s(U)\, \mathrm{d}x \approx \tfrac{1}{2}h\big(S_{i-1} + S_i\big), \tag{4.47}$$

for $i = 1, \ldots, J$ where $S_i = S(U_i)$. Then (4.46) can be written as

$$\begin{pmatrix} 1 & 0 & \ldots & 0 \\ -1 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \\ \vdots & & -1 & 1 \end{pmatrix} \begin{pmatrix} U_1 \\ \vdots \\ \vdots \\ U_J \end{pmatrix} = \tfrac{1}{2}h \begin{pmatrix} S_0 + S_1 \\ S_1 + S_2 \\ \vdots \\ S_{J-1} + S_J \end{pmatrix} + \begin{pmatrix} u_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4.48}$$

We can see that (4.48) gives the following finite difference approximation to (4.34):

$$\frac{U_{j+1} - U_j}{h} = \tfrac{1}{2}\big(S_j + S_{j+1}\big), \tag{4.49}$$

for $j = 0, \ldots, J-1$ with $U_0 = u_0$ as the boundary condition.. This gives a simple finite difference scheme for the ODE (4.34). This has been derived by applying a P-G method to a piecewise linear approximation U of the solution u and using the trapezoidal rule to evaluate the integrals. We now wish to apply this procedure to a time dependent problem to obtain the box scheme.

4.3.2 The mildly nonlinear time dependent problem

Let us first consider the following mildly nonlinear partial differential equation on $x \in (0,1)$ and $t \in (0,1)$ with a source term:

$$u_t + u_x = S(u). \tag{4.50}$$

This can be written as $Lu = S$ where $Lu := u_t + u_x$ with the appropriate initial and boundary conditions. We now seek a piecewise bilinear approximation U to u which depends on x and t, i.e.

$$U(x,t) = \sum_{n=0}^{N} \sum_{j=0}^{J} U_j^n \phi_j(x) \phi_n(t). \tag{4.51}$$

Following the procedure described above the P-G method requires that the residual is orthogonal to the test functions χ_i^k defined as

$$\chi_i^k(x,t) = \begin{cases} 1 & \text{if } (x,t) \in (x_{i-1}, x_i) \times (t_{k-1}, t_k) \\ 0 & \text{otherwise,} \end{cases} \tag{4.52}$$

for $i = 1, \ldots, J$ and $k = 1, \ldots, N$. So

$$\langle LU - S, \chi_i^k(x,t) \rangle = 0, \tag{4.53}$$

where $\langle \cdot, \cdot \rangle$ is now a double integral

$$\langle u, v \rangle = \int_0^1 \int_0^1 u(x,t) v(x,t) \mathrm{d}x \, \mathrm{d}t. \tag{4.54}$$

Then (4.53) becomes

$$\sum_{n=0}^{N} \sum_{j=0}^{J} U_j^n \langle \phi_j \phi_n', \chi_i^k \rangle + \sum_{n=0}^{N} \sum_{j=0}^{J} U_j^n \langle \phi_j' \phi_n, \chi_i^k \rangle = \langle S, \chi_i^k \rangle. \tag{4.55}$$

for $i = 1, \ldots, J$ and $k = 1, \ldots, N$ (using linearity of the l^2 inner product to bring the inner product inside the summation). Now

$$\langle S, \chi_i^k \rangle = \int_0^1 \int_0^1 S(U)\chi_i^k(x,t)\mathrm{d}x\,\mathrm{d}t = \int_{t_{k-1}}^{t_k} \int_{x_{i-1}}^{x_i} S(U)\mathrm{d}x\,\mathrm{d}t =: I_i^k, \tag{4.56}$$

for $i = 1, \ldots, J$ and $k = 1, \ldots, N$. Also

$$\langle \phi_j \phi_n', \chi_i^k \rangle = \int_{t_{k-1}}^{t_k} \phi_n'(t)\mathrm{d}t \int_{x_{i-1}}^{x_i} \phi_j(x)\mathrm{d}x,$$

and

$$\langle \phi_j' \phi_n, \chi_i^k \rangle = \int_{t_{k-1}}^{t_k} \phi_n(t)\mathrm{d}t \int_{x_{i-1}}^{x_i} \phi_j'(x)\mathrm{d}x,$$

where the integrals are obtained from (4.19) and (4.43). We can easily see that the general form for each entry in the matrix system (4.55) is given by

$$\tfrac{1}{2}\triangle x\Big[-(U_j^n + U_{j+1}^n) + (U_j^{n+1} + U_{j+1}^{n+1})\Big] + \tfrac{1}{2}\triangle t\Big[-(U_j^n - U_{j+1}^n) - (U_j^{n+1} - U_{j+1}^{n+1})\Big] = I_{j+1}^{n+1}, \tag{4.57}$$

for $j = 0, \ldots, J-1$ and $n = 0, \ldots, N-1$. Let us approximate I_{j+1}^{n+1} by the trapezoidal rule, i.e.

$$\begin{aligned} I_{j+1}^{n+1} &= \int_{t_n}^{t_{n+1}} \int_{x_j}^{x_{j+1}} S(U)\mathrm{d}x\,\mathrm{d}t \\ &= \tfrac{1}{4}\triangle x \triangle t \left[S_j^n + S_j^{n+1} + S_{j+1}^n + S_{j+1}^{n+1}\right], \end{aligned} \tag{4.58}$$

where $S_j^n := S(U_j^n)$. Then, dividing (4.57) by $\triangle x \triangle t$, gives

$$\begin{aligned} &\tfrac{1}{2\triangle t}\left[\left(U_{j+1}^{n+1} - U_{j+1}^n\right) + \left(U_j^{n+1} - U_j^n\right)\right] + \tfrac{1}{2\triangle x}\left[\left(U_{j+1}^{n+1} - U_j^{n+1}\right) + \left(U_{j+1}^n - U_j^n\right)\right] \\ &\qquad = \left[S_j^n + S_j^{n+1} + S_{j+1}^n + S_{j+1}^{n+1}\right], \end{aligned} \tag{4.59}$$

which is precisely the box scheme applied to (4.50). We have interpreted the box scheme as a P-G method using the trapezoidal rule to evaluate the integrals. Moreover, we have presented its derivation in detail so that it is easily generalised to a mesh that is non-uniform in x, or non-uniform in t, or both. Indeed, the only requirement is that the trial spaces are spanned by tensor product basis functions. Let us now consider a nonlinear flux term.

4.3.3 The nonlinear time dependent problem

Suppose we wish to find the piecewise linear approximation to the solution of the following nonlinear conservation law:

$$Lu := u_t + f(u)_x = 0, \tag{4.60}$$

with appropriate initial and boundary conditions. If the piecewise linear approximation to the solution of (4.60) is defined by (4.51) then the P-G method requires that

$$\langle LU(x,t), \chi_i^k(x,t)\rangle = 0, \tag{4.61}$$

for $i = 1,\ldots,J$ and $k = 1,\ldots,N$ where χ_i^k is defined by (4.52). This becomes

$$\int_{t_{k-1}}^{t_k}\int_{x_{i-1}}^{x_i} (U_t + f(U)_x)\mathrm{d}x\,\mathrm{d}t = 0. \tag{4.62}$$

Let D denote the rectangle $D = \{(x,t)|\ (x,t) \in (x_{i-1},x_i)\times(t_{k-1},t_k)\}$. Then using the Divergence Theorem (2.126) (from Chapter 2) this becomes

$$\int_{x_{i-1}}^{x_i} \big[-U(x,t_{k-1}) + U(x,t_k)\big]\mathrm{d}x + \int_{t_{k-1}}^{t_k} \big[f(U(x_i,t)) - f(U(x_{i-1},t))\big]\mathrm{d}t = 0. \tag{4.63}$$

Let us approximate the second term by the trapezoidal rule, i.e.

$$\begin{aligned}\int_{t_{k-1}}^{t_k} \big[f(U(x_i,t)) - f(U(x_{i-1},t))\big]\mathrm{d}t &= \tfrac{1}{2}\triangle t\big[(F_i^k + F_i^{k-1}) - (F_{i-1}^k + F_{i-1}^{k-1})\big] \\ &=: \tfrac{1}{2}\triangle t Q_i^k,\end{aligned} \tag{4.64}$$

where, in the usual way, $F_i^k = f(U_i^k)$. Also

$$\int_{x_{i-1}}^{x_i} \big[U(x,t_k) - U(x,t_{k-1})\big]\mathrm{d}x = \sum_{n=0}^{N}\sum_{j=0}^{J} U_j^n\big(\phi_n(t_k) - \phi_n(t_{k-1})\big)\int_{x_{i-1}}^{x_i}\phi_j(x)\mathrm{d}x.$$

Now, since

$$\phi_n(t_k) - \phi_n(t_{k-1}) = \begin{cases} 1 & n = k \\ -1 & n = k-1 \\ 0 & \text{otherwise,}\end{cases}$$

and

$$\int_{x_{i-1}}^{x_i}\phi_j(x)\mathrm{d}x = \begin{cases} \frac{1}{2}\triangle x & j = i \\ \frac{1}{2}\triangle x & j = i-1 \\ 0 & \text{otherwise,}\end{cases}$$

we obtain the following matrix system to solve for U_j^n for $j = 0, \ldots, J$ and $n = 0, \ldots, N$:

$$\begin{pmatrix} -1 & 1 & 0 & \ldots \\ 0 & \ddots & \ddots & \\ \vdots & & -1 & 1 \end{pmatrix} \begin{pmatrix} U_0^0 & \ldots & U_0^N \\ \vdots & & \vdots \\ U_J^0 & \ldots & U_J^N \end{pmatrix} \begin{pmatrix} 1 & 0 & \ldots \\ 1 & \ddots & \\ 0 & \ddots & 1 \\ \vdots & & 1 \end{pmatrix} + \nu \begin{pmatrix} Q_1^1 & \ldots & Q_1^N \\ \vdots & & \vdots \\ Q_J^1 & \ldots & Q_J^N \end{pmatrix} = 0.$$

The general form for each entry is then

$$\Big[\big(U_{j+1}^{n+1} - U_{j+1}^n\big) + \big(U_j^{n+1} - U_j^n\big)\Big] + \nu\Big[\big(F_{j+1}^{n+1} - F_j^{n+1}\big) + \big(F_{j+1}^n - F_j^n\big)\Big] = 0, \quad (4.65)$$

for $j = 0, \ldots, J-1$ and $n = 0, \ldots, N-1$. This is precisely the box scheme applied to (4.60). Again, the derivation shows how it is easily generalised to a non-uniform mesh.

4.4 Modifying the box scheme for shocks

The box scheme is inadequate for computing discontinuous solutions of nonlinear conservation laws. As discussed in Chapter 11 of (LeVeque 1992), it is expected that the numerical method will have difficulty near the discontinuity. It is typically found that first order methods give very smeared solutions while second order methods give oscillations. The box scheme is a second order method and, if we again examine Figure 4-1, we can see oscillations forming around the discontinuity. With the viewpoint of Section 4.3, where we derived the box scheme as a P-G method, we can develop alternatives. We begin by looking at the nonlinear conservation law when the shock is already formed, i.e. the Riemann problem (4.5) with initial data (4.6). Initially we restrict attention to the case when $u_l > u_r \geq 0$ and take as our test case the most famous model problem, Burgers' equation

$$u_t + \left(\tfrac{1}{2}u^2\right)_x = 0. \quad (4.66)$$

The shock speed is simply $\frac{1}{2}(u_l + u_r)$ and so we know the position of the shock at each time level. We can use this information to construct an algorithm to eliminate the oscillations observed when applying the basic box scheme (without any correction to take account of the shock). The box scheme sweeps from left to right at each level n for all j because we prescribe data on the left boundary (see (4.7)).

$$u(0,t) = u_l, \quad (4.67)$$

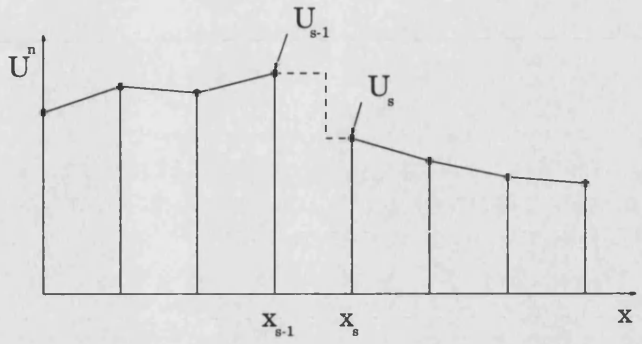

Figure 4-2: *A diagram showing the shock location within cell* (x_k, x_{k+1}) *at time level* t^n. *We are assuming that the trial space is piecewise constant either side of the shock.*

for all $0 \leq t \leq T$. In Chapter 2 we derived the basic box scheme applied to a linear conservation law $u_t + au_x = 0$. The same can be done for (4.5) by integrating it over the cell $(x_j, x_{j+1}) \times (t^n, t^{n+1})$ to obtain

$$\int_{x_j}^{x_{j+1}} \left[u(x, t^{n+1}) - u(x, t^n)\right] \mathrm{d}x + \int_{t^n}^{t^{n+1}} \left[f(u(x_{j+1}, t)) - f(u(x_j, t))\right] \mathrm{d}t = 0. \quad (4.68)$$

Approximating these integrals by the trapezium rule gives precisely the box scheme (4.65).

The equation (4.65) is the basic box scheme which we will use as a starting point for our algorithm. We wish to apply this numerical scheme in some way across a cell which contains the shock. This is where we will modify the box scheme to take into account the discontinuity. In Section 4.3 we derived the box scheme as a P-G method using a piecewise constant test space and a piecewise linear trial space. In the cell which contains the shock we now use a trial space which is piecewise constant but augmented by a shock, i.e. piecewise constant either side of the shock within the cell.

Figure 4-2 shows a diagram of the numerical scheme at level n. The shock occurs in cell (x_{s-1}, x_s) and, since the trial space is assumed to be piecewise constant either side of the shock, we need one extra piece of information to determine the extra parameter at each time level. This is the simplest extension of the trial space at t^n and we call the extra parameter the *shock position*. We actually need to know both the location of the cell which contains the shock and the position of the shock within that cell. For the moment we assume that the shock has already formed and so the location of the cell is given for each n. We will have to split the trial space to integrate the conservation law around the whole cell (taking into account the discontinuity) and across the shock. An added difficulty is that the shock might have moved to the next cell when going from time level n to $n + 1$. We consider this case first.

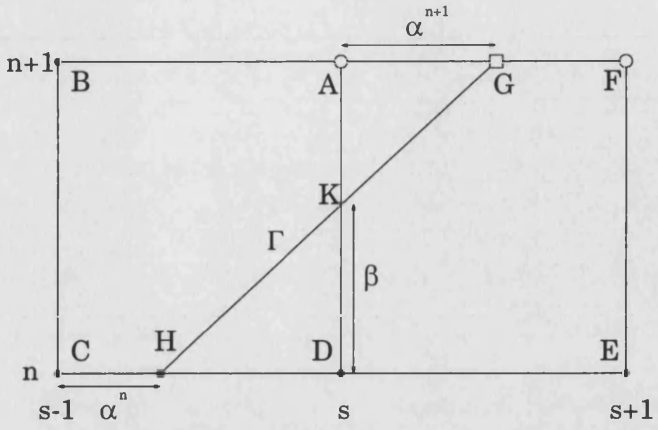

Figure 4-3: *The case when the shock moves to the next cell at the new time level.*

4.4.1 Double cell analysis for existing shock

Suppose the shock occurs in the cell (x_{s-1}, x_s) at t^n and in the cell (x_s, x_{s+1}) at t^{n+1}, as shown in Figure 4-3. We assume that the shock is given by a straight line, denoted by Γ, and crosses the $x = x_s$ line at the point $t = t^n + \beta\triangle t$. We have three unknowns to find: α^{n+1}, U_s^{n+1} and U_{s+1}^{n+1}.

In the two cells which contain the shock we use a trial space which is piecewise constant but augmented by a shock, i.e. piecewise constant either side of the shock within the appropriate cell. Hence in the left cell this occurs at level n and in the right cell at $n+1$. So, to find our unknowns, we must integrate the conservation law around the two boxes $(x_{s-1}, x_s) \times (t^n, t^{n+1})$ and $(x_s, x_{s+1}) \times (t^n, t^{n+1})$ shown in Figure 4-3 using the new trial space. Equation (4.68) still holds but we now use this updated trial space to calculate the integrals in the x direction at level n in the left cell and at level $n+1$ in the right. For example, in the left cell U^n is constant either side of the shock and so we integrate $U^n(x)$ exactly, i.e.

$$\int_{x_{s-1}}^{x_s} u(x, t^n)\,\mathrm{d}x = \triangle x\big[\alpha^n u(x_{s-1}, t^n) + (1-\alpha^n)u(x_s, t^n)\big].$$

The integrals in the t direction on the far left and far right remain the same but the integral across the centre (i.e the boundary between the two boxes) has to take into account the shock. For now we denote this as $F_s^{n+1/2}$ and discuss how to evaluate it below. Hence, conservation over the left box gives

$$\tfrac{1}{2}\big(U_{s-1}^{n+1} + U_s^{n+1}\big) - \big[\alpha^n U_{s-1}^n + (1-\alpha^n)U_s^n\big] + \nu\Big[F_s^{n+1/2} - \tfrac{1}{2}\big(F_{s-1}^n + F_{s-1}^{n+1}\big)\Big] = 0. \tag{4.69}$$

Similarly in the box $(x_s, x_{s+1}) \times (t^n, t^{n+1})$ we have

$$\left[\alpha^{n+1}U_s^{n+1} + (1-\alpha^{n+1})U_{s+1}^{n+1}\right] - \tfrac{1}{2}(U_s^n + U_{s+1}^n) + \nu\left[\tfrac{1}{2}(F_{s+1}^n + F_{s+1}^{n+1}) - F_s^{n+1/2}\right] = 0. \tag{4.70}$$

The trial space is piecewise constant either side of the shock and so, at level n, U_{s-1}^n is the value to the left of the shock and U_s^n is the value to the right; similarly, at level $n+1$, U_s^{n+1} and U_{s+1}^{n+1} are the values to the left and right respectively.

From Figure 4-3 we can see that the shock divides the two cells into four regions: two pentagons and two triangles. Consider the bilinear approximation in the left pentagon ABCHK. Along the right-hand edge (KA) there is a uniform linear variation in t. In the top triangle AKG, the variation along the left side of the shock (KG) is obtained from turning this vertical variation along KA into the shock-direction variation. Similarly, the vertical variation also gives the variation along the shock edge defining the pentagon (HK). Hence it is the same linear variation along the whole shock trajectory HG. This means we can integrate the conservation law (4.5) around a parallelogram that shrinks onto the whole shock length HG to obtain

$$\begin{aligned}&\left\{\tfrac{1}{2}(\alpha^{n+1}+1-\alpha^n)(U_s^{n+1}+U_{s-1}^n) - \tfrac{1}{2}\nu\left[F_s^{n+1}+F_{s-1}^n\right]\right\}\\ &-\left\{\tfrac{1}{2}(\alpha^{n+1}+1-\alpha^n)(U_{s+1}^{n+1}+U_s^n) - \tfrac{1}{2}\nu\left[F_{s+1}^{n+1}+F_s^n\right]\right\} = 0.\end{aligned} \tag{4.71}$$

Also, the previously unspecified quantity $F_s^{n+1/2}$ is given by

$$F_s^{n+1/2} = \tfrac{1}{2}\beta\left[\beta F_{s+1}^{n+1} + (2-\beta)F_s^n\right] + \tfrac{1}{2}(1-\beta)\left[(1+\beta)F_s^{n+1} + (1-\beta)F_{s-1}^n\right], \tag{4.72}$$

where, since Γ is a straight line

$$\beta = \frac{1-\alpha^n}{1-\alpha^n+\alpha^{n+1}}. \tag{4.73}$$

4.4.2 Single cell analysis for existing shock

If Δt is small enough then the shock may stay in the same cell at the new time level (i.e. occur in cell (x_{s-1}, x_s) at both t^n and t^{n+1}). In this case we are in a situation as shown in Figure 4-4. We still have three unknowns (α^{n+1}, U_s^{n+1} and U_{s+1}^{n+1}) but U_{s+1}^{n+1} can be calculated using the basic box scheme. We now need to find the two unknowns in the left cell. Conservation around this cell gives

$$\begin{aligned}&\left[\alpha^{n+1}U_{s-1}^{n+1} + (1-\alpha^{n+1})U_s^{n+1}\right] - \left[\alpha^n U_{s-1}^n + (1-\alpha^n)U_s^n\right]\\ &\quad+\tfrac{1}{2}\nu\left[(F_s^n + F_s^{n+1}) - (F_{s-1}^n + F_{s-1}^{n+1})\right] = 0,\end{aligned} \tag{4.74}$$

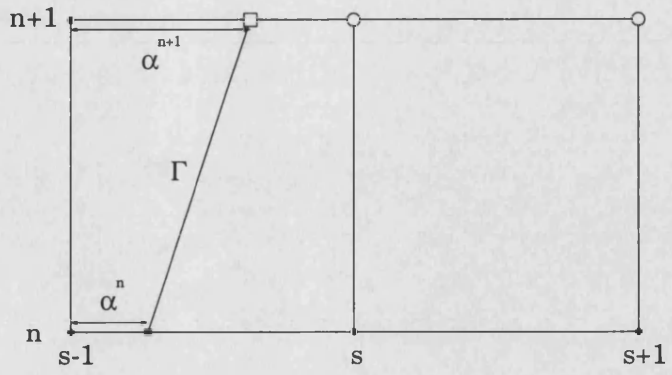

Figure 4-4: *The case when the shock stays in the same cell at the new time level.*

In this case, at level n, U_{s-1}^n is the value to the left of the shock and U_s^n is the value to the right; similarly, at level $n+1$, U_{s-1}^{n+1} and U_s^{n+1} are the values to the left and right respectively. Hence, as in the double cell case, integrating the conservation law (4.5) around a parallelogram that shrinks onto the shock gives

$$\begin{aligned}&\left\{\tfrac{1}{2}(\alpha^{n+1}-\alpha^n)\left(U_{s-1}^{n+1}+U_{s-1}^n\right)-\tfrac{1}{2}\nu\left[F_{s-1}^{n+1}+F_{s-1}^n\right]\right\}\\&-\left\{\tfrac{1}{2}(\alpha^{n+1}-\alpha^n)\left(U_s^{n+1}+U_s^n\right)-\tfrac{1}{2}\nu\left[F_s^{n+1}+F_s^n\right]\right\}=0.\end{aligned}\tag{4.75}$$

Once we have solved these two equations to find the unknowns α^{n+1} and U_s^{n+1} we can use the basic box scheme to find U_j^{n+1} for $j=s+1,\ldots,J$.

4.4.3 Shift from the double cell to the single cell

Suppose we are in the double cell case and integrate the conservation law over both cells in Figure 4-3. Information is always known along the left side, along the bottom and along half the right side (the lower half). Let us consider the flux along the top and half the upper half of the right side, and denote this by $F_D^{\neg}$, i.e.

$$F_D^{\neg}=\tfrac{1}{2}\left(U_{s-1}^{n+1}+U_s^{n+1}\right)+\alpha^{n+1}U_s^{n+1}+(1-\alpha^{n+1})U_{s+1}^{n+1}+\tfrac{1}{2}\nu F_{s+1}^{n+1}.\tag{4.76}$$

Now let $\alpha^{n+1}\to 0$. Then the shock is at the node $j=s$ at the new time level and we suppose that $U_s^{n+1}\to U_{s-}^{n+1}$. Then

$$F_D^{\neg}\longrightarrow\tfrac{1}{2}\left(U_{s-1}^{n+1}+U_{s-}^{n+1}\right)+U_{s+1}^{n+1}+\tfrac{1}{2}\nu F_{s+1}^{n+1},\quad\text{as}\quad\alpha^{n+1}\longrightarrow 0.\tag{4.77}$$

We can follow the same procedure for the single cell case and integrate the conservation law over both cells in Figure 4-4. Then, the flux along the top and the upper half of

the right side, which we denote $F_S^{\neg}$, is

$$F_S^{\neg} = \alpha^{n+1}U_{s-1}^{n+1} + (1-\alpha^{n+1})U_s^{n+1} + \tfrac{1}{2}(U_s^{n+1} + U_{s+1}^{n+1}) + \tfrac{1}{2}\nu F_{j+1}^{n+1}. \tag{4.78}$$

If we let $\alpha^{n+1} \to 1$ then the shock is again at the node $j = s$ at the new time level and we now suppose $U_s^{n+1} \to U_{s+}^{n+1}$. Then

$$F_S^{\neg} \longrightarrow U_{s-1}^{n+1} + \tfrac{1}{2}(U_{s+}^{n+1} + U_{s+1}^{n+1}) + \tfrac{1}{2}\nu F_{s+1}^{n+1}, \quad \text{as} \quad \alpha^{n+1} \longrightarrow 1. \tag{4.79}$$

The limits in (4.77) and (4.79) are equal if

$$\tfrac{1}{2}(U_{s-1}^{n+1} + U_{s-}^{n+1}) + U_{s+1}^{n+1} + \tfrac{1}{2}\nu F_{s+1}^{n+1} = U_{s-1}^{n+1} + \tfrac{1}{2}(U_{s+}^{n+1} + U_{s+1}^{n+1}) + \tfrac{1}{2}\nu F_{s+1}^{n+1},$$

which reduces to

$$U_{s+}^{n+1} - U_{s-}^{n+1} = U_{s+1}^{n+1} - U_{s-1}^{n+1}. \tag{4.80}$$

This is called the *shock jump*. We will need this relation if, in the iteration based on the double cell, we find that $\alpha^{n+1} < 0$ (and so the shock is really in the same cell at the new time level). Then, the trial space in the left cell will have to be altered to include the shock at level $n+1$. When we assumed the shock had moved cells this was piecewise linear and so, to make sure the equations are consistent, as we cross the $j = s$ boundary we must change U_s^{n+1} by the shock jump. Newton's method can be applied to equations (4.74) and (4.75) to solve for α^{n+1} and U_s^{n+1} with appropriate starting values. These are obtained from the values previously calculated in the double cell case and so we take $1 + \alpha^{n+1}$ and $U_s^{n+1} + U_{s+}^{n+1} - U_{s-}^{n+1}$ as the starting values of α^{n+1} and U_s^{n+1} respectively. In practice the latter is found by using the shock jump condition (4.80) and so becomes $U_s^{n+1} + U_{s+1}^{n+1} - U_{s-1}^{n+1}$.

4.4.4 Description of the overall algorithm

To implement this algorithm we need to set up the data on the boundaries. In discrete form (4.6) is simply

$$U_j^0 = \begin{cases} u_l, & x_j < \sigma \\ u_r, & x_j > \sigma, \end{cases} \tag{4.81}$$

for $j = 0, \ldots, J$ and (4.67) becomes $U_0^n = u_l$ for $n = 0, \ldots, N$. We first find the index $j = s-1$ which denotes the location of the shock at level $n = 0$ and then set

$$\alpha^0 = \frac{\sigma - x_{s-1}}{\Delta x}. \tag{4.82}$$

We assume that σ is not a nodal value and so $\alpha^0 \in (0,1)$. For each level n we find U_{j+1}^{n+1} by solving

$$\tfrac{1}{2}\nu(U_{j+1}^{n+1})^2+U_{j+1}^{n+1}+(1-\tfrac{1}{2}\nu U_j^{n+1})U_j^{n+1}-(1-\tfrac{1}{2}\nu U_{j+1}^n)U_{j+1}^n-(1+\tfrac{1}{2}\nu U_j^n)U_j^n=0, \quad (4.83)$$

for $j = 0,\ldots,s-2$. This is the basic box scheme for Burgers' equation (i.e. (4.65) with $F_j^n := f(U_j^n) = \frac{1}{2}(U_j^n)^2$), which is solved to the left of the shock.

We are now at the point where we must modify the box scheme to allow for the shock. Assume the cell goes into the next cell at the new time level. Then, the three nonlinear equations (4.69), (4.70) and (4.71) must be solved, and we use Newton's method to do this. Hence we need to define our starting guesses for the unknowns α^{n+1}, U_s^{n+1} and U_{s+1}^{n+1}. The shock speed at level n can be approximated as $\frac{1}{2}(U_{s-1}^n + U_s^n)$ (i.e. the shock speed for Burgers' equation, see (LeVeque 1992, page 29)). This suggests that a reasonable starting guess for α^{n+1} is $-(1-\alpha^n)+\frac{1}{2}(U_{s-1}^n+U_s^n)\nu$. Starting guesses for U_s^{n+1} and U_{s+1}^{n+1} are given by U_{s-1}^n and U_s^n respectively.

Hence the algorithm is as follows:

1. solve (4.83) for $j = 0,\ldots,s-2$;
2. iterate to find U_s^{n+1}, U_{s+1}^{n+1} and α^{n+1} by solving (4.69), (4.70) and (4.71) with starting guesses U_{s-1}^n, U_s^n and $-(1-\alpha^n)+\frac{1}{2}(U_{s-1}^n+U_s^n)\nu$, respectively;
3. if $\alpha^{n+1} > 0$ find U_{j+1}^{n+1} for $j = s+1,\ldots,J-1$ using (4.83). Set $s = s+1$ (since the shock is now in the next cell) and move to the next time level (i.e. go to step 1.). Otherwise, change U_s^{n+1} by the shock jump defined in (4.80), i.e. set

$$\bar{U}_s^{n+1} = U_s^{n+1} + U_{s+1}^{n+1} - U_{s-1}^{n+1}, \quad (4.84)$$

where U_s^{n+1} and U_{s+1}^{n+1} are the values calculated in step 1. Since α^{n+1} is negative, also set

$$\bar{\alpha}^{n+1} = 1 + \alpha^{n+1}, \quad (4.85)$$

where α^{n+1} is the value calculated in step 1. Go to step 4.;
4. solve (4.74) and (4.75) using Newton's method to re-calculate α^{n+1} and U_s^{n+1}. Take $\bar{\alpha}^{n+1}$ and $\bar{U}_s^{n+1}$ as starting guesses. Use (4.83) to re-calculate U_{s+1}^{n+1} and then to find U_{j+1}^{n+1} for $j = s+1,\ldots,J-1$. Move to the next time level (i.e. go to step 1.).

This is called the **corrected box scheme**. Note that in solving the quadratic equation in (4.83) to find U_{j+1}^{n+1} we take the positive root. This is because the positive root can

be shown to satisfy

$$\left(U_{j+1}^{n+1}\right)_+ = -U_j^{n+1} + U_{j+1}^n + U_j^n + O(\nu) \longrightarrow U_{j+1}^n - \left(U_j^{n+1} - U_j^n\right), \quad \text{as} \quad \nu \longrightarrow 0,$$

whereas

$$\left(U_{j+1}^{n+1}\right)_- = -\frac{2}{\nu} - \left(-U_j^{n+1} + U_{j+1}^n + U_j^n\right) + O(\nu) \longrightarrow -\infty, \quad \text{as} \quad \nu \longrightarrow 0.$$

4.4.5 Alternative calculation of α^{n+1} for existing shock

The algorithm described above has to solve either two or three coupled nonlinear equations to find the unknowns α^{n+1}, U_s^{n+1} and U_{s+1}^{n+1} (depending on whether the shock has moved to the next cell). If, at each time level, we were able to estimate α^{n+1} in some way then the nonlinear computations would be much simpler.

For Burgers' equation with Riemann data we know α^{n+1} since the shock is already formed at $n = 0$. This is given by $\alpha^{n+1} = \alpha^n + \frac{1}{2}(u_l + u_r)\nu$. If this value was used then numerical results can show that the algorithm gives the exact solution. However, as we will discuss below, the aim is to solve a nonlinear conservation law for the situation when the initial data does not involve a shock but will eventually go into a shock at a certain time (which we will henceforth call *shock-forming data*). In this case α^{n+1} is not known and will have to be estimated.

From our analysis of the P-G method in Section 4.2.1, on finding a piecewise linear approximation of a discontinuous function of the form

$$u(x_i) = \begin{cases} u_l, & i = 0, \ldots, k-1 \\ u_r, & i = k, \ldots, J, \end{cases} \tag{4.86}$$

the resulting coefficients are observed to have an oscillatory structure about u_r (see (4.31)). We can rearrange the first oscillatory value to obtain an expression for the shock proportion α

$$\alpha = \frac{U_k + u_l - 2u_r}{2(u_l - u_r)}.$$

We could use this value of α as an estimation for α^{n+1} by setting $U_k = U_s^{n+1}$ (we need to take this at level $n+1$ to be consistent with the shock proportion). However, U_s^{n+1} has not yet been calculated and so this also needs to be estimated. Suppose we take the usual trial space for the box scheme in the cell (x_{s-1}, x_s) except along the edge where $t = t^n$. The shock lies along this edge and so we use a piecewise constant trial space augmented by the shock. The equation to solve is simply (4.74) with $\alpha^{n+1} = \frac{1}{2}$.

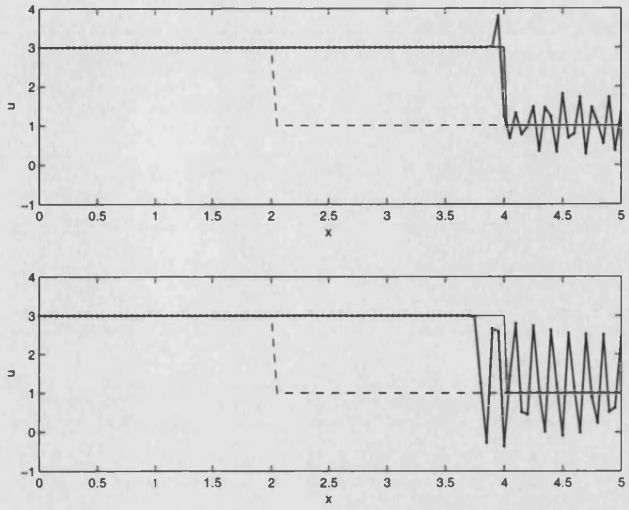

Figure 4-5: *Burgers' equation with Riemann data. The top plot shows the box scheme with the P-G estimation for* α^{n+1} *and the bottom plot shows the box scheme with no modification. The numerical solution is given by dots joined by an unbroken line, the exact solution by a thin unbroken line and the initial condition by a dashed line. The results are shown at time* $t = 1$ *and* $\nu = 0.25$.

This will give an estimate for U_s^{n+1}, which we will denote $\tilde{U}_s^{n+1}$. Then we set

$$\alpha^{n+1} = \frac{\tilde{U}_s^{n+1} + u_l - 2u_r}{2(u_l - u_r)}. \tag{4.87}$$

We call this the *P-G estimation* for α^{n+1}. Once this is found we follow a similar procedure to the algorithm above. Figure 4-5 shows the result of the algorithm using the P-G estimation (in the top plot) compared with the basic box scheme without any modification (in the bottom plot). However, the P-G estimation still leads to oscillations but it is an improvement on the basic box scheme. The theory developed in Section 4.2.1 showed that projecting discontinuities onto a piecewise linear trial space gives oscillations which we can then use to deduce information about the shock position. We tried to utilise this here but have shown that it is not very effective. Also, it is very limiting since this estimation can only really be applied to the shock Riemann problem.

4.4.6 Numerical results

The corrected box scheme is now compared with the exact solution (4.8) and other well known finite difference schemes used for Burgers' equation with Riemann data. Let us take $u_l = 3$, $u_r = 1$, $\sigma = 2$ and assume $0 \leq x \leq 5$. Figure 4-6 shows plots of the following schemes: the upwind method, Lax-Wendroff and MacCormack methods,

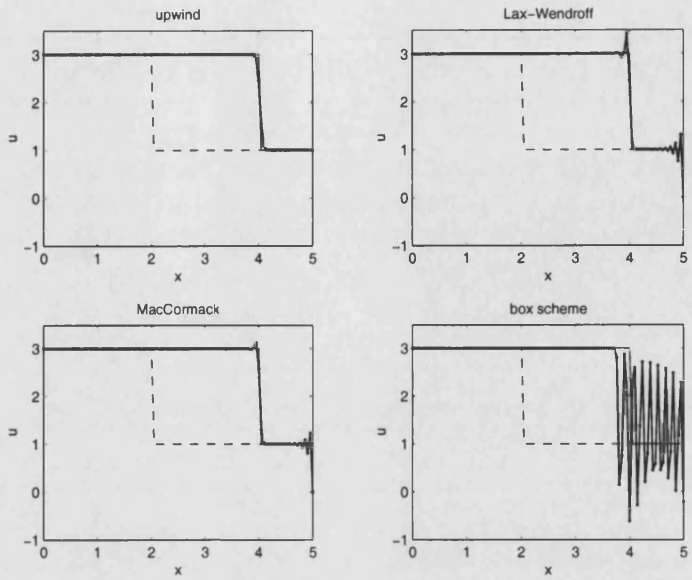

Figure 4-6: *Plots of four finite difference schemes applied to Burgers' equation with Riemann data compared with the exact solution for* $\triangle x = 5/105$ *and* $\triangle t = 1/84$ *at fixed* $t = 1$. *The numerical solution is shown as dots joined by an unbroken line, the exact solution as a thin unbroken line and the initial condition as a dashed line. The schemes are: upwind (top left), Lax-Wendroff (top right), MacCormack (bottom left) and the box scheme (bottom right).*

(LeVeque 1992, page 127), and the box scheme. In these plots $\nu = 0.25$ (with $\triangle t = 1/84$ and $\triangle x = 5/105$, which is chosen so σ is not a nodal value) and the solution is plotted at $t = 1$. We see that all the second order methods give oscillations and the box scheme is the most oscillatory.

In Chapter 3 we managed to reduce (and even eliminate) the oscillations by using the weighted-box scheme. In Figure 4-7 the first three plots show the solution for three values of $\theta > \frac{1}{2}$. The oscillations can be significantly reduced but this introduces smearing and we aim to eliminate them completely. The bottom right plot shows the corrected box scheme applied to the same problem. The oscillations have completely disappeared and the solution is moving at the correct speed. This is very reassuring although using Riemann data is only a first step: we would like the corrected box scheme to be accurate and free of oscillations for more complicated initial conditions. Firstly, we could consider initial data which already has a shock but is not piecewise constant either side, e.g.

$$u(x,0) = \begin{cases} -\frac{1}{2}x + 4, & x < 2 \\ \frac{1}{3}\left(-\frac{1}{2}x + 4\right), & x > 2, \end{cases} \tag{4.88}$$

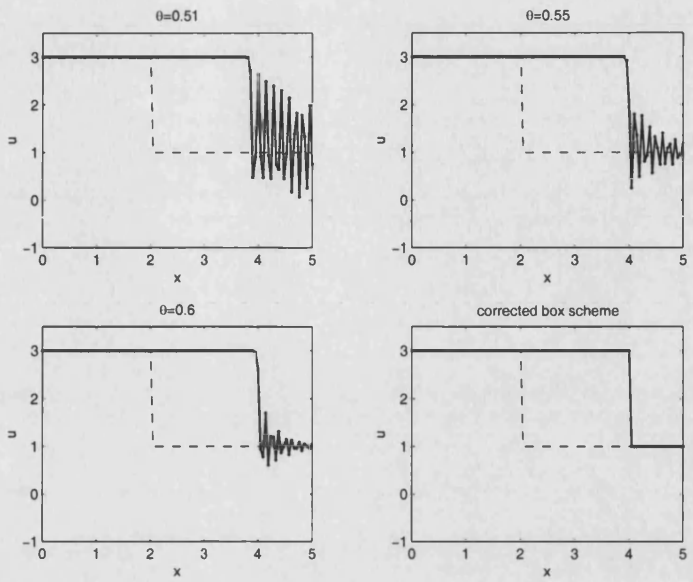

Figure 4-7: *The first three plots show the weighted box scheme (with* $\theta = 0.51$, 0.55 *and* 0.6*) applied to Burgers' equation with Riemann data compared with the exact solution for* $\triangle x = 5/105$ *and* $\triangle t = 1/84$ *at fixed* $t = 1$*. The bottom right plot shows the corrected box scheme. The numerical solution is shown as dots joined by an unbroken line, the exact solution as a thin unbroken line and the initial condition as a dashed line.*

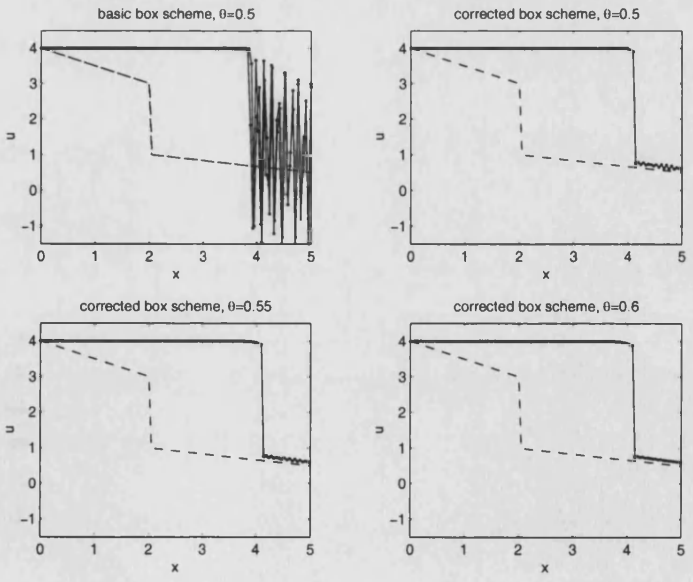

Figure 4-8: *Numerical solution (shown as dots joined by an unbroken line) of Burgers' equation with initial data given by (4.88) compared with the exact solution for* $\triangle x = 5/105$ *and* $\triangle t = 1/84$ *at fixed* $t = 1$*. The top left plot shows the basic box scheme and top right shows the corrected box scheme. The bottom two plots show the corrected weighted box scheme.*

We now apply the corrected box scheme to this problem and the result is shown in the top right plot in Figure 4-8. This is a great improvement on the basic box scheme (top left plot) although the oscillations have not disappeared completely. We have also considered incorporating θ into the corrected box scheme and the results are shown in the bottom two plots in Figure 4-8. The oscillations have reduced as θ is increased from $\frac{1}{2}$ and so we can improve the scheme further. There is a modification of the discretised equations in the algorithm described above to include the weighting. This is outlined now.

The integration of the fluxes along the left and right edges of the box (see (4.68)) now uses a weighting rather than the usual trapezoidal rule. So, the basic box scheme applied to Burgers' equation, which is given by (4.83), now becomes

$$\nu\theta U_{j+1}^{n+1^2}+U_{j+1}^{n+1}+\left(1-\nu\theta U_j^{n+1}\right)U_k^{n+1}-\left[1-\nu(1-\theta)U_{j+1}^n\right]U_{j+1}^n-\left[1+\nu(1-\theta)U_j^n\right]U_j^n=0. \tag{4.89}$$

This equation is used for all j either side of the shock. We now have to modify the equations in the shock cell (or cells) to incorporate θ. In the double cell case (4.69) and (4.70) become

$$\begin{aligned}&\tfrac{1}{2}\left(U_{s-1}^{n+1}+U_s^{n+1}\right)-\left[\alpha^n U_{s-1}^n-(1-\alpha^n)U_s^n\right]\\&+\nu\left\{F_s^{n+1/2}-\left[(1-\theta)F_{s-1}^n+\theta F_{s-1}^{n+1}\right]\right\}=0,\end{aligned} \tag{4.90}$$

and

$$\begin{aligned}&\left[\alpha^{n+1}U_s^{n+1}+(1-\alpha^{n+1})U_{s+1}^{n+1}\right]-\tfrac{1}{2}\left(U_s^n-U_{s+1}^n\right)\\&+\nu\left\{\left[(1-\theta)F_{s+1}^n+\theta F_{s+1}^{n+1}\right]-F_s^{n+1/2}\right\}=0,\end{aligned} \tag{4.91}$$

whilst everything else remains unchanged. In the single cell case (4.74) becomes

$$\begin{aligned}&\left[\alpha^{n+1}U_{s-1}^{n+1}+(1-\alpha^{n+1})U_s^{n+1}\right]-\left[\alpha^n U_{s-1}^n-(1-\alpha^n)U_s^n\right]\\&+\nu\left\{\left[(1-\theta)F_s^n+\theta F_s^{n+1}\right]-\left[(1-\theta)F_{s-1}^n+\theta F_{s-1}^{n+1}\right]\right\}=0.\end{aligned} \tag{4.92}$$

In both cases we do not weight the shock relations (4.71) and (4.75).

4.4.7 The shock problem with $u_r < 0$

We now consider the shock Riemann problem with $u_r < 0$ but still assuming the shock propagates from left to right, and so $u_l + u_r > 0$. In this case data is prescribed on the right boundary and so $u(X,t) = u_r$ for all $0 \le t \le T$. We now have one less unknown but the same number of equations. When data is known on the right boundary we will solve from the right to the left to find the unknown value U_j^{n+1} (rather than U_{j+1}^{n+1}

when going from left to right). Hence for Burgers' equation the basic box scheme from right to left involves the following quadratic for U_j^{n+1}:

$$\tfrac{1}{2}\nu(U_j^{n+1})^2 - U_j^{n+1} - \left(1 + \tfrac{1}{2}\nu U_{j+1}^{n+1}\right) U_{j+1}^{n+1} + \left(1 - \tfrac{1}{2}\nu U_{j+1}^{n}\right) U_{j+1}^{n} + \left(1 + \tfrac{1}{2}\nu U_j^{n}\right) U_j^{n} = 0, \tag{4.93}$$

When solved, we take the negative square root since

$$\left(U_j^{n+1}\right)_- = -U_{j+1}^{n+1} + U_{j+1}^{n} + U_j^{n} + O(\nu) \longrightarrow U_{j+1}^{n} - \left(U_{j+1}^{n+1} - U_j^{n}\right), \quad \text{as} \quad \nu \longrightarrow 0,$$

whereas

$$\left(U_j^{n+1}\right)_+ = \frac{2}{\nu} + \left(-U_{j+1}^{n+1} + U_{j+1}^{n} + U_j^{n}\right) + O(\nu) \longrightarrow +\infty, \quad \text{as} \quad \nu \longrightarrow 0.$$

For each level n we begin by using (4.83) to solve from the left and (4.93) to solve from the right. At the node j, where the solution crosses from being positive to negative, either of these equations could be used to find the unknown value. Hence we have an over-determined system. To remedy this situation we note that conservation will not be satisfied at this cross-over point. Hence we need to integrate over the two cells which contain the node j. Then

$$\int_{x_{j-1}}^{x_{j+1}} \left[u(x, t^{n+1}) - u(x, t^n)\right] \mathrm{d}x + \int_{t_n}^{t_{n+1}} \left[f(u(x_{j+1}, t)) - f(u(x_{j-1}, t))\right] \mathrm{d}t = 0. \tag{4.94}$$

These integrals are now approximated by the trapezoidal rule. This leads to an explicit formula for U_j^{n+1} in terms of its five neighbouring values (it is explicit because there is no integral of the flux at $x = x_j$), thus

$$\begin{aligned} U_j^{n+1} &= U_j^n + \tfrac{1}{2}\left(\left[U_{j+1}^n + U_{j-1}^n\right] - \left[U_{j+1}^{n+1} + U_{j-1}^{n+1}\right]\right) \\ &\quad + \tfrac{1}{2}\nu\left(\left[F_{j-1}^{n+1} + F_{j-1}^n\right] - \left[F_{j+1}^{n+1} + F_{j+1}^n\right]\right). \end{aligned} \tag{4.95}$$

We now describe the algorithm in this case.

As in Section 4.4.4 we set up the initial and boundary data (with the addition that we define $U_J^n = u_r$ for all $n = 0, \ldots, N$) and find α^0 and the cell (x_{s-1}, x_s) where the shock is located. We then solve the basic box scheme to the left and right of the shock using (4.83) for $j = 0, \ldots, s-2$ and (4.93) for $j = J-1, \ldots, s+1$.

Figure 4-9 shows a schematic diagram of corrected box scheme across the shock. We first assume that the shock has moved to the next cell (to the right) at the new time level. In this case there are only two unknowns: α^{n+1} and U_s^{n+1} (unlike the situation for $u_l \geq 0$ where U_{s+1}^{n+1} is also unknown). We use the same shock condition as for the double cell analysis in the $u_r \geq 0$ case (i.e. equation (4.71)). This must be coupled with

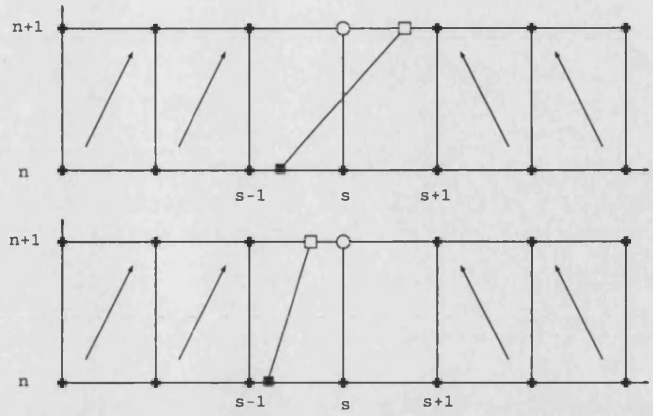

Figure 4-9: *A schematic diagram for the corrected box scheme when* $u_r < 0$ *(and* $u_l + u_r > 0$*). In the top figure the shock has moved to the next cell at the new time level and in the bottom figure the shock is in the same cell.*

the equation which describes conservation over the two cells (x_{s-1}, x_s) and (x_s, x_{s+1}). The relation in (4.94) still holds but the trial space has now changed to take into account the position of the shock. Hence we obtain

$$\begin{aligned}\alpha^n U^n_{s-1} + \left(\tfrac{3}{2} - \alpha^n\right) U^n_s - \left(\tfrac{1}{2} + \alpha^{n+1}\right) U^{n+1}_s - (1 - \alpha^{n+1}) U^{n+1}_{s+1} + \tfrac{1}{2}\left(U^n_{s+1} - U^{n+1}_{s-1}\right) \\ + \tfrac{1}{2}\nu\left[F^{n+1}_{s-1} + F^n_{s-1}\right] - \tfrac{1}{2}\nu\left[F^{n+1}_{s+1} + F^n_{s+1}\right] = 0. \qquad (4.96)\end{aligned}$$

We solve (4.71) and (4.96) using Newton's method to find the unknowns α^{n+1} and U^{n+1}_s with starting guesses $-(1 - \alpha^n) + \frac{1}{2}\left(U^n_{s-1} + U^n_s\right)\nu$ and U^n_{s-1} respectively. If $\alpha^{n+1} > 0$ then we set $s = s + 1$ and move onto the next time level. Otherwise, the shock has stayed in the same cell at the new time level. We follow the same procedure as described in step 2. of the algorithm in Section 4.4.4. The (already) calculated U^{n+1}_s is changed by the shock jump and we add one to α^{n+1}. These are taken to be the starting guesses in the Newton loop. The equations to solve are the shock relation (4.75) and the following equation describing conservation over the two cells (again taking into account the position of the shock):

$$\begin{aligned}\alpha^n U^n_{s-1} + \left(\tfrac{3}{2} - \alpha^n\right) U^n_s - \left(\tfrac{3}{2} - \alpha^{n+1}\right) U^{n+1}_s - \alpha^{n+1} U^{n+1}_{s-1} + \tfrac{1}{2}\left(U^n_{s+1} - U^{n+1}_{s+1}\right) \\ + \tfrac{1}{2}\nu\left[F^{n+1}_{s-1} + F^n_{s-1}\right] - \tfrac{1}{2}\nu\left[F^{n+1}_{s+1} + F^n_{s+1}\right] = 0. \qquad (4.97)\end{aligned}$$

Then proceed to the next time level.

The numerical results are shown in Figure 4-10. The top two plots show the box scheme without any modification to take into account the shock for two initial conditions (the first is Riemann and the second piecewise linear either side of the shock). We only have

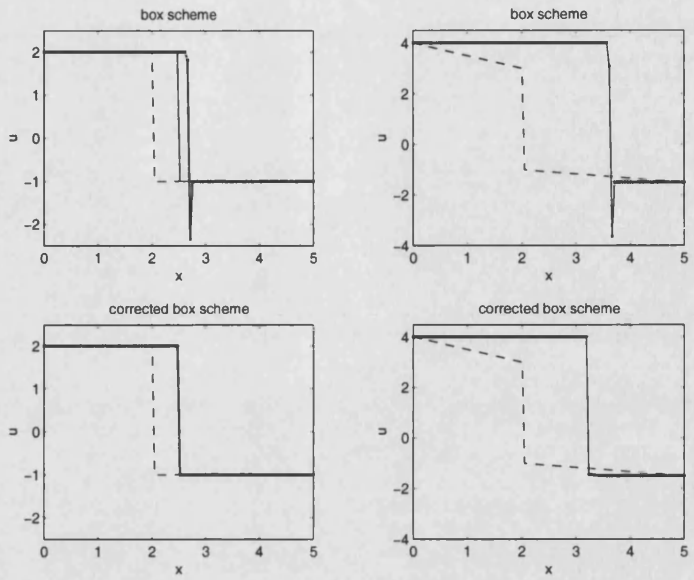

Figure 4-10: *Plots of the box scheme (top two plots) compared with the corrected box scheme (bottom two plots) for Burgers' equation with negative initial data (shown as dots joined by an unbroken line)* $\triangle x = 5/105$ *and* $\triangle t = 1/84$ *at fixed* $t = 1$. *The left two plots use Riemann data (*$u_l = 2$ *and* $u_r = -1$*) and so the exact solution is shown as a thin unbroken line.*

the exact solution for the Riemann case but this clearly shows that the shock moves at the wrong speed (note that there are no oscillations in this case). The bottom two plots shows how the corrected box scheme shock moves at the right speed. Hence the algorithm is very effective when $u_r < 0$ for initial data already in shock form.

4.5 Shock-forming data

In this section we extend the algorithm described in Section 4.4 and apply it to Burgers' equation with two initial conditions which will become a shock solution after some finite time. The first has two piecewise constant sections joined linearly, i.e.

$$u(x,0) = \begin{cases} 3, & 0 \leq x \leq 1 \\ -2x + 5, & 1 < x < 2 \\ 1, & x \geq 2, \end{cases} \tag{4.98}$$

and the second is a smooth curve given by

$$u(x,0) = 2 + \tanh(9 - 5x). \tag{4.99}$$

For both examples we set $u(0,t) = u(0,0)$ for all t. The exact solution of Burgers' equation with (4.98) is given by

$$u(x,t) = \begin{cases} 3, & 0 \leq x \leq 1+3t \\ \frac{-2x+5}{1-2t}, & 1+3t < x < 2+t \\ 1, & x \geq 2+t, \end{cases} \tag{4.100}$$

for $t < \frac{1}{2}$ and

$$u(x,t) = \begin{cases} 3, & x < \frac{1}{2}(3+4t) \\ 1, & x > \frac{1}{2}(3+4t), \end{cases} \tag{4.101}$$

for $t \geq \frac{1}{2}$. If the four finite difference schemes considered in the previous sections are applied to Burgers' equation, with either (4.98) and (4.99), then the same conclusions can be made: the basic box scheme is the most oscillatory and these oscillations cannot be eliminated by increasing θ from $\frac{1}{2}$. This is provided t is chosen so that the shock has already formed. We now apply the corrected weighted box scheme.

We wish to use the weighted box scheme without modification until the time when the shock occurs and then apply the algorithm. So, a procedure must be found to predict when the shock has formed. For the example (4.98) we know this happens at $t = \frac{1}{2}$ (from the exact solution). However, this is a special case and in general the exact solution will not be known (as for (4.99)). From the theory of shock formation in conservation laws we can find the breaking time T_b (i.e. the time when shock forms) by considering the solution written in terms of the initial data, $u_0(x)$. We have to restrict $u_0(x)$ to be smooth with $u_0'(x)$ somewhere negative. For Burgers' equation the breaking time is simply (LeVeque 1992, page 25)

$$T_b = -\frac{1}{\min\left[u_0'(x)\right]}. \tag{4.102}$$

and $T_b = 1/5$ for (4.99). Once we have found the breaking time we can find the value m where we will start to use the corrected box scheme (i.e level $n = m$ corresponds to the initial level in the algorithm). However, we first need to find the shock location at that level. Define

$$\bar{U}_j^n = \left|U_{j+1}^n - U_j^n\right|, \tag{4.103}$$

for $j = 0, \ldots, J-1$ and $n = 0, \ldots, N$. Then find the index which gives

$$\max_{0 \leq j \leq J-1} \bar{U}_j^m,$$

and label this as index l. Then the shock occurs in the cell (x_{l-1}, x_l) to $O(\triangle x)$ and so we set $s = l$. The only thing left to define is the shock proportion at the initial level (i.e. to specify α^m). However, we are assuming that no shock has formed at this level

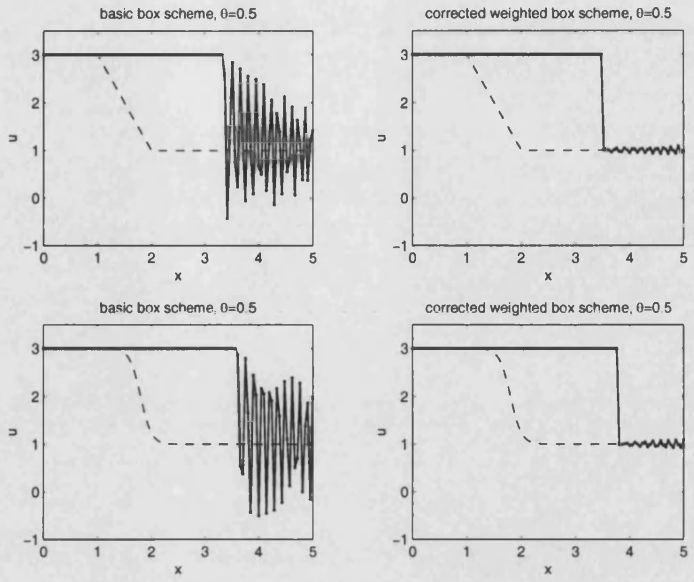

Figure 4-11: *The left two plots show the basic weighted box scheme (with $\theta = \frac{1}{2}$) applied to Burgers' equation with initial data given by (4.98) and (4.99) for $\triangle x = 5/105$ and $\triangle t = 1/84$ at fixed $t = 1$. The right two plots show the corrected weighted box scheme in these two cases.*

and so set $\alpha^m = \frac{1}{2}$. The algorithm is now applied for $n = m, \ldots, N-1$. The right two plots in Figure 4-11 show the results of applying this algorithm for the two examples (4.98) and (4.99). We can compare this with the basic box scheme which is shown in the left plots. The corrected weighted box scheme greatly improves the solution and, although not reproduced here, as θ is increased from $\frac{1}{2}$ they disappear completely.

However, the procedure for finding the level n where the shock first occurs is only valid for Burgers' equation (and smooth initial data). We need a more general, reliable way of estimating this which can be applied to any nonlinear conservation law. Suppose we write $u_t + f(u)_x = 0$ in the form

$$u_t + a(u)u_x = 0, \tag{4.104}$$

where $a(u) = f'(u)$. We know that the shock forms when the characteristics start to cross. In terms of the numerical solution this is equivalent to saying

$$\left[a(U_j^n) - a(U_{j+1}^n)\right]\triangle t > \xi \triangle x, \tag{4.105}$$

for $j = 0, \ldots, J-1$ where ξ is a constant to be determined. We start by solving the basic box scheme and at each level n calculate

$$S_j^n = \left[a(U_j^n) - a(U_{j+1}^n)\right]\nu, \tag{4.106}$$

for $j = 0, \ldots, J-1$. We also store the maximum value of S_j^n, i.e.

$$E^n = \max_{0 \le j \le J-1} S_j^n, \tag{4.107}$$

and the index l^n which denotes the cell where this maximum occurs for each level n. We continue until $E^n > \xi$. This is the time at which the shock occurs and so we set $n = m$. We now find the cell where the shock occurs by setting $s = l^m$, then choose $\alpha^m = 0.5$ and apply the algorithm for $n = m, \ldots, N-1$.

The only parameter still unknown is ξ. After some numerical experiments we find that choosing $\xi \approx 0.2$ gives the correct location of the shock. It can lead to an underestimation of the shock location which results in starting the algorithm too early. However, the scheme seems robust enough to cope with this. Experiments can show that using this value of ξ gives very similar results to example from Figure 4-11. The same experiment was tried with $\xi = 0.25$ and $\xi = 0.3$: the former case showed no difference in the solution but there were significantly more oscillations in the latter case (which is to be expected as the algorithm is then started too late). Hence the algorithm works well if ξ is chosen around 0.2 and it does not matter if it is started too early.

Lastly, we make a comment about how the algorithm behaves immediately after it has been implemented. Consider the example in Figure 4-11 where the results are shown at $t = 1$. For the piecewise linear data (4.98) the shock occurs at $t = \frac{1}{2}$ and so Figure 4-12 shows the results plotted at both $t = \frac{1}{2}$ and $t = 1$ when $\xi = 0.2$ is used to guess the position of the shock. The algorithm has just been implemented at $t = \frac{1}{2}$ since using ξ slightly underestimates the location of the shock. The oscillations are quite severe but this is to be expected since an oscillatory solution was used as the input to the algorithm and $\theta = \frac{1}{2}$. This Figure demonstrates the effectiveness of the algorithm since, by $t = 1$, the oscillations are severely reduced. Hence our algorithm is very robust: even if oscillatory data is given at the input level, it still manages to damp the oscillations as time progresses. The parameter ξ will need to be estimated for different fluxes as it is directly related to the shock speed; this is a drawback but ξ is relatively easy to find.

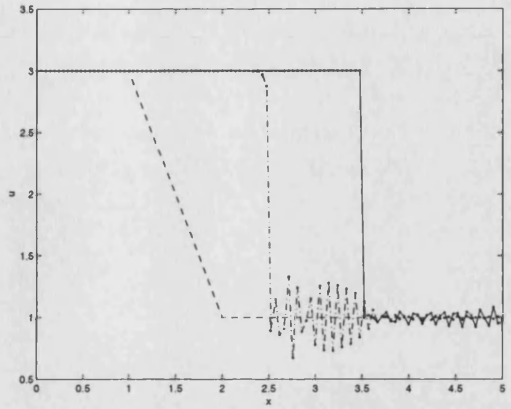

Figure 4-12: *Plot of the corrected box scheme applied to Burgers' equation with initial data given by (4.98) for* $\triangle x = 5/105$ *and* $\triangle t = 1/84$ *with* $\theta = 0.5$. *The dots joined by an unbroken line line denotes the numerical solution at* $t = 1$ *and the dots joined by a dot-dashed line denotes the numerical solution at* $t = 0.5$.

4.6 The Langmuir Model

At the beginning of this Chapter we stated that the motivation behind studying nonlinear conservation laws was because, in physical applications, a coupled system involving a conservation law and reaction equation can be approximated by an Equilibrium model which has a nonlinear flux term. An example of this is the Langmuir Model, where the Equilibrium model is given by (4.4). This can also be written as

$$\big(a + g(a)\big)_t + V a_x = 0, \tag{4.108}$$

where $g(a) = B\lambda a/(\mu + \lambda a)$. In this case the flux term is connected to the t derivative so we could think of interchanging x and t to apply the algorithm described in Section 4.4. Assuming Riemann data is defined on the $x = 0$ axis, i.e.

$$a(0,t) = \begin{cases} a_l, & t < \tau \\ a_r, & t > \tau, \end{cases} \tag{4.109}$$

for any τ, we can find the shock speed, s. This is given by

$$s = \frac{\lambda\mu B + (\mu + \lambda a_l)(\mu + \lambda a_r)}{V(\mu + \lambda a_l)(\mu + \lambda a_r)}. \tag{4.110}$$

However, as shown for the Linear Model, the Equilibrium model does not accurately describe the full dynamics of the problem. We would like to consider the Langmuir Model in the form of (4.1) and (4.2). It is not clear whether the solution will give a shock wave but we can study the travelling wave solution to see how the wave profile

propagates.

4.6.1 The travelling wave solution

Consider the Langmuir Model as defined in (4.1) and (4.2). Following a procedure described in (Grindrod 1991, pages 34-43) and (Whitham 1974, pages 101-102) we seek a solution in the form of a travelling wave. That is, we define

$$a(x,t) = \bar{a}(z), \qquad b(x,t) = \bar{b}(z), \tag{4.111}$$

where $z = x - Ut$. Here U is the wave speed which to be determined. Then (4.1) and (4.2) become

$$(U-V)\bar{a}_z - \lambda\bar{a}(B-\bar{b}) + \mu\bar{b} = 0 \tag{4.112}$$

$$U\bar{b}_z + \lambda\bar{a}(B-\bar{b}) - \mu\bar{b} = 0. \tag{4.113}$$

Adding (4.112) and (4.113) eliminates the source term and so

$$(U-V)\bar{a}_z + U\bar{b}_z = 0. \tag{4.114}$$

Suppose

$$\{\bar{a};\ \bar{b}\} \longrightarrow \{a_l, a_r; b_l, b_r\}, \quad \text{as} \quad z \longrightarrow \mp\infty \tag{4.115}$$

with

$$\lambda a_l(B-b_l) = \mu b_l, \qquad \lambda a_r(B-b_r) = \mu b_r, \tag{4.116}$$

to ensure the system possesses a solution commensurate with the boundary conditions. Then, on integrating (4.114) from $-\infty$ to z we can use (4.116) to obtain an expression for b entirely in terms of a

$$\bar{b} = \frac{\lambda B a_l}{\mu + \lambda a_l} - \frac{U-V}{U}(\bar{a} - a_l). \tag{4.117}$$

Hence (4.112) becomes

$$(U-V)\bar{a}_z - \lambda\bar{a}B + (\mu+\lambda\bar{a})\left[\frac{\lambda B a_l}{\mu+\lambda a_l} - \frac{U-V}{U}(\bar{a}-a_l)\right] = 0. \tag{4.118}$$

This differential equation can be used to find the travelling wave speed U. Assuming

$$\bar{a}_z \longrightarrow 0 \quad \text{as} \quad z \longrightarrow \mp\infty, \tag{4.119}$$

we can take the limit of (4.118) as $z \to +\infty$ (the left hand side of (4.118) is automatically satisfied as $z \to -\infty$). So

$$\lambda a_r B + (\mu + \lambda a_r)\left[\frac{\lambda B a_l}{\mu + \lambda a_l} - \frac{U - V}{U}(a_r - a_l)\right] = 0,$$

which can be rearranged to give

$$U = \frac{V(\mu + \lambda a_l)(\mu + \lambda a_r)}{\lambda\mu B + (\mu + \lambda a_l)(\mu + \lambda a_r)}. \tag{4.120}$$

Comparing this with the shock speed in (4.110) for the Equilibrium model shows that $U = 1/s$. Hence the travelling wave and Equilibrium model speeds are identical (since the x and t axes have been interchanged in (4.108)). Using the expression for U in (4.120), we can write the ODE in (4.118) as

$$(\bar{a} - a_r)(a_l - \bar{a}) = -2\beta\bar{a}_z, \tag{4.121}$$

where

$$\beta := \frac{V(\mu + \lambda a_l)(\mu + \lambda a_r)}{2\lambda\left[\lambda\mu B + (\mu + \lambda a_l)(\mu + \lambda a_r)\right]} = \frac{U}{2\lambda}. \tag{4.122}$$

This has solution

$$\bar{a} = a_r + \frac{a_l - a_r}{1 + \exp\left(\frac{a_l - a_r}{2\beta}z\right)}, \tag{4.123}$$

and so $\bar{a}(z) \to a_l$ as $z \to -\infty$ and $\bar{a}(z) \to a_l$ as $z \to +\infty$, which agrees with the conditions for $\bar{a}$ in (4.115). This means there is a shift from a_l to a_r as it propagates. Finally, the travelling wave solution of the Langmuir Model is

$$\bar{a} = a_r + \frac{a_l - a_r}{1 + \exp\left(\frac{a_l - a_r}{U/\lambda}(x - Ut)\right)}. \tag{4.124}$$

It is interesting to note that this analysis is very similar to an investigation of the travelling wave solution of the viscous Burgers' equation carried out by (Whitham 1974, 101-102). This is the simplest equation which combines both nonlinear propagation effects and diffusive effects. Consider

$$u_t + uu_x = \nu u_{xx}. \tag{4.125}$$

The travelling wave solution $\bar{u}$ satisfies

$$-U\bar{u}_z + \bar{u}\bar{u}_z = \nu\bar{u}_{zz}, \tag{4.126}$$

which can be integrated to give

$$-U\bar{u} + \tfrac{1}{2}\bar{u}^2 + C = \nu \bar{u}_z,$$

where C is an arbitrary constant. Assuming $u \to u_l$, u_r and $u_z \to 0$ as $z \to \mp\infty$, it follows that $U = \frac{1}{2}(u_l + u_r)$ and $C = \frac{1}{2}u_l u_r$. Hence the solution of (4.126) is

$$\bar{u} = u_r + \frac{u_l - u_r}{1 + \exp\left(\frac{u_l - u_r}{2\nu}(x - Ut)\right)}. \tag{4.127}$$

As $\nu \to 0$ this smooth solution converges to the shock solution. In (Whitham 1974) this travelling wave solution is compared to the exact solution of (4.126) (with Riemann data). It is shown that Riemann data diffuses into the steady profile (4.127) as $t \to \infty$.

4.6.2 Numerical experiments

In this section we observe some of the phenomena discussed in the analysis above by plotting the Langmuir Model against t for a range of values of λ and μ. In the travelling wave solution for inviscid Burgers' equation we saw that as $\nu \to 0$ the smooth solution converges to the shock solution. This is equivalent to requiring $\lambda \to \infty$ for the smooth travelling wave solution of the Langmuir Model to converge to a shock (we can see this directly by comparing (4.124) and (4.127)). Hence as λ gets larger we should see the shift from a_l to a_r becoming steeper. Although the solution will never go into a shock, (4.124) shows that when λx is large the wave front will be very steep. It is in this case that wish to apply the algorithm from Section 4.4. Let us fix x and plot the numerical solution of the Langmuir Model against t. Assume the initial condition for a is zero (and similarly for b) and the boundary condition is piecewise linear pulse given by

$$a(x,0) = a_l, \qquad a(0,t) = \begin{cases} 1, & t < 1 \\ 2t - 1, & 1 \le t \le 2 \\ 3, & t > 2, \end{cases} \tag{4.128}$$

Then $a_l = 1$ and $a_r = 3$ in the terminology of the previous section. We now have $a_l < a_r$ because we assume data is prescribed on the t boundary (to be consistent with the work from previous Chapters). Assume that b satisfies the equilibrium condition on both boundaries, i.e.

$$b(0,t) = \frac{B\lambda a(0,t)}{\mu + \lambda a(0,t)}, \qquad b(x,0) = \frac{B\lambda a(x,0)}{\mu + \lambda a(x,0)}. \tag{4.129}$$

Figure 4-13 shows the box-trap scheme applied to the Langmuir Model with this data at $x = 10$ (chosen so that λx is relatively large). In the top two plots $\lambda = \mu = 1$ and in the bottom two plots $\lambda = 3$, $\mu = 1$. The left two plots show that, at the same

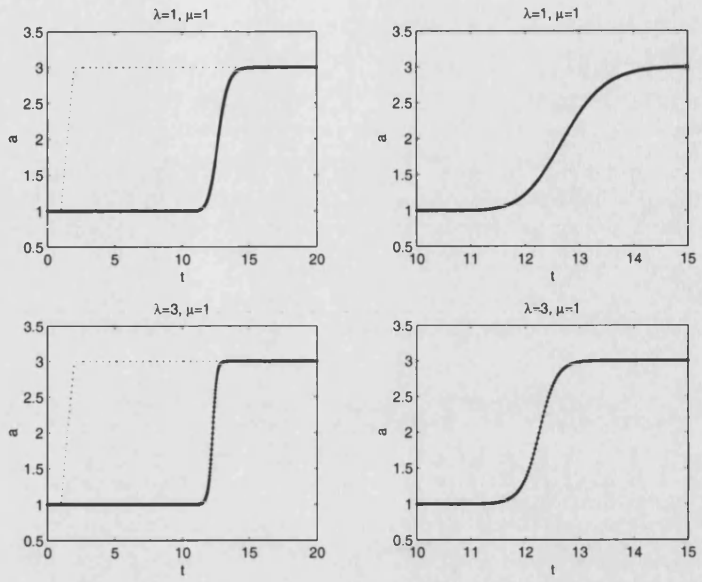

Figure 4-13: *Plot of the box-trap scheme against t (shown as dots joined by an unbroken line) applied to the Langmuir model at a fixed* $x = 10$ *and* $V = 1$. *The dashed line denotes the boundary condition. In the top two plots* $\lambda = \mu = 1$ *and in the bottom two plots* $\lambda = 3$, $\mu = 1$.

distance x, the larger value of λ gives a steeper wave front. In the right two plots we have enlarged the region where the shift occurs from a_l to a_r. When $\lambda = 1$ the shift roughly takes place between $t = 11.5$ and $t = 14.5$ and when $\lambda = 3$, between $t = 12$ and $t = 13$. Hence the wave has become steeper by a factor of λ which is what we expect from the analysis of the travelling wave solution (4.124).

Now consider piecewise constant boundary data, i.e.

$$a(x,0) = 0, \qquad a(0,t) = \begin{cases} a_l, & t < t^* \\ a_r, & t > t^*, \end{cases} \tag{4.130}$$

We use non-smooth boundary data to really test the weighted box-trap scheme: this is the situation where there will be most oscillations. In Figures 4-14 and 4-15 we plot four values of λ and μ and change θ to see whether the oscillations are eliminated. Figure 4-14 shows results for $\lambda = 3$, $\mu = 1$ and $\lambda = 9$, $\mu = 1$: in the left two plots $\theta = 0.5$ and in the right two plots $\theta = 0.5 + \triangle t$. In both cases the very oscillatory part of the numerical solution disappears. However, as λ becomes very large, as shown in Figure 4-15, the oscillations have not disappeared completely by increasing θ. We could try increasing θ further but this will introduce diffusion into the numerical scheme. Instead we will correct the box scheme around the steep front.

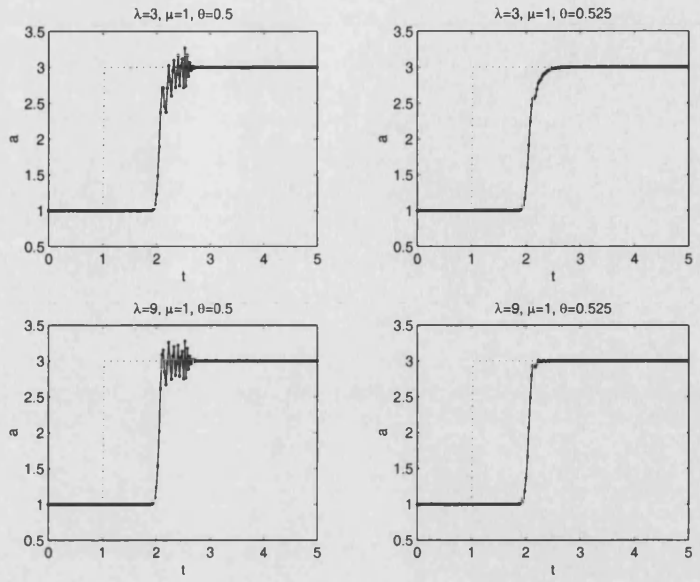

Figure 4-14: *Plot of the weighted box-trap scheme against t (shown as dots joined by an unbroken line) applied to the Langmuir model at a fixed $x = 1$ and $V = 1$. The dashed line denotes the boundary condition. In the top two plots $\lambda = 3$, $\mu = 1$ (with $\theta = 0.5$ in the left and $\theta = 0.5 + \triangle t = 0.525$ in the right) and in the bottom two plots $\lambda = 9$, $\mu = 1$ (with $\theta = 0.5$ in the left and $\theta = 0.5 + \triangle t = 0.525$ in the right).*

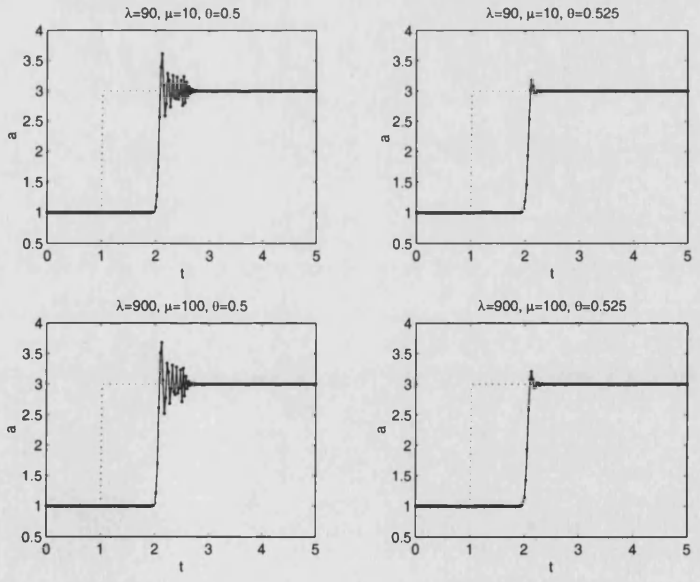

Figure 4-15: *Plot of the weighted box-trap scheme against t (shown as dots joined by an unbroken line) applied to the Langmuir model at a fixed $x = 1$ and $V = 1$. The dashed line denotes the boundary condition. In the top two plots $\lambda = 90$, $\mu = 10$ (with $\theta = 0.5$ in the left and $\theta = 0.5 + \triangle t = 0.525$ in the right) and in the bottom two plots $\lambda = 900$, $\mu = 100$ (with $\theta = 0.5$ in the left and $\theta = 0.5 + \triangle t = 0.525$ in the right).*

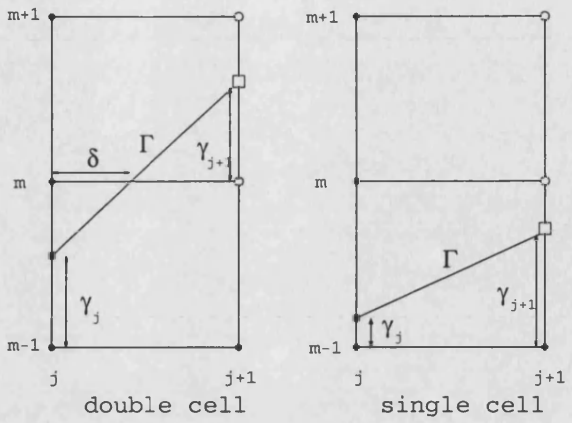

Figure 4-16: *A diagram showing the shock position when it either moves to the next cell (left picture) or stays in the same cell (right picture) at the next spatial level.*

4.6.3 The corrected weighted box-trap scheme

Suppose (4.1) is treated as a nonlinear conservation law (i.e. where a is a nonlinear function of c) and so we write the Langmuir Model in terms of a and c as follows:

$$c_t + V a_x = 0 \tag{4.131}$$

$$c_t - a_t = \lambda a(B - c + a) - \mu(c - a). \tag{4.132}$$

The initial and boundary data for c are found using the data for a and b in (4.130) and (4.129). We apply the weighted box-trap scheme to (4.131) and (4.132) with this data. In general, for the box scheme, it does not matter whether n is fixed and the discretised equations are solved for all j or the reverse procedure is applied. However, since the shock jump appears on the t boundary it makes sense to fix j, solve for every n, and then move onto the next spatial step. Consider the diagram in Figure 4-16. The basic discretised equations for (4.131) and (4.132) are given by

$$C_{j+1}^{n+1} - C_{j+1}^{n} + C_j^{n+1} - C_j^n + 2p\theta\big(A_{j+1}^{n+1} - A_j^{n+1}\big) + 2p(1-\theta)\big(A_{j+1}^n - A_j^n\big) = 0, \tag{4.133}$$

and

$$\begin{aligned}\big(C_{j+1}^{n+1} - A_{j+1}^{n+1}\big) - \big(C_{j+1}^n - A_{j+1}^n\big) &= \tfrac{1}{2}\lambda' A_{j+1}^{n+1}\big(B - C_{j+1}^{n+1} + A_{j+1}^{n+1}\big) - \tfrac{1}{2}\mu'\big(C_{j+1}^{n+1} - A_{j+1}^{n+1}\big) \\ &\quad + \tfrac{1}{2}\lambda' A_{j+1}^{n}\big(B - C_{j+1}^{n} + A_{j+1}^{n}\big) - \tfrac{1}{2}\mu'\big(C_{j+1}^{n} - A_{j+1}^{n}\big).\end{aligned} \tag{4.134}$$

In Appendix C the algorithm is described which solves the weighted box-trap scheme when a shock profile is applied on the left boundary. It is very similar to the algorithm

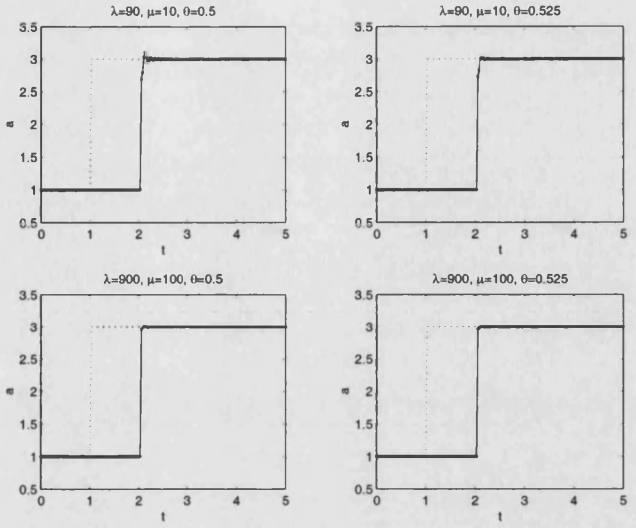

Figure 4-17: *Plot of the corrected weighted box-trap scheme against t (shown as dots joined by an unbroken line) applied to the Langmuir model at a fixed* $x = 1$ *and* $V = 1$. *The dashed line denotes the boundary condition. In the top two plots* $\lambda = 90$, $\mu = 10$ *(with* $\theta = 0.5$ *in the left and* $\theta = 0.5 + \triangle t = 0.525$ *in the right) and in the bottom two plots* $\lambda = 900$, $\mu = 100$ *(with* $\theta = 0.5$ *in the left and* $\theta = 0.5 + \triangle t = 0.525$ *in the right).*

in Section 4.4 except we now have more unknowns (five if the shock moves to the next cell and three if the shock stays in the same cell) as the Langmuir Model involves a pair of equations to solve.

In the double cell case the weighted box-trap scheme is applied in both the bottom and top cells separately, taking into account the position of the shock. The five unknowns are A_{j+1}^{m+1}, C_{j+1}^{m+1}, A_{j+1}^{m}, C_{j+1}^{m} and γ_{j+1}. In the top cell (4.132) has to be solved using the trapezoidal scheme at level $j+1$. Although the shock lies along this edge of the cell, we do not take account of it in the trapezoidal scheme (it is only when applying the box scheme to (4.131) that this modification occurs). Similarly, in the single cell case (with the three unknowns being A_{j+1}^{m}, C_{j+1}^{m} and γ_{j+1}) we just apply the trapezoidal scheme to solve (4.132) across the right edge of the box at level $j+1$.

The corrected weighted box-trap scheme is applied to the same problem as considered in Figure 4-15. The results are shown in Figure 4-17. There is a great reduction in the oscillations. When $\lambda = 90$ and $\mu = 10$ the weighting θ still needs to be larger than $\frac{1}{2}$ but as these increase we see that $\theta = \frac{1}{2}$ is sufficient to remove all the oscillations.

Chapter 5

The Flushing-through Model

5.1 Introduction and derivation

In this Chapter we analyse a more realistic model to describe the transport of chemicals in groundwater flow. It involves several chemical species and therefore gives a system of coupled partial differential equations. Section 1.3.3 of the Introduction discussed the motivation behind this area of research and mentioned a system of six partial differential equations with quadratic nonlinearities that could be used as a basis to analyse larger systems. This was initially posed by AEA Technology Harwell (now SERCO Assurance) and is given by

$$\begin{aligned}
(a_1)_t + V(a_1)_x - \epsilon(a_1)_{xx} &= -\lambda_1 a_1 + \mu_1 b_1 - \gamma a_1 a_2 + \delta a_3 && (5.1)\\
(b_1)_t &= \lambda_1 a_1 - \mu_1 b_1 && (5.2)\\
(a_2)_t + V(a_2)_x - \epsilon(a_2)_{xx} &= -\lambda_2 a_2 + \mu_2 b_2 - \gamma a_1 a_2 + \delta a_3 && (5.3)\\
(b_2)_t &= \lambda_2 a_2 - \mu_2 b_2 && (5.4)\\
(a_3)_t + V(a_3)_x - \epsilon(a_3)_{xx} &= -\lambda_3 a_3 + \mu_3 b_3 + \gamma a_1 a_2 - \delta a_3 && (5.5)\\
(b_3)_t &= \lambda_3 a_3 - \mu_3 b_3, && (5.6)
\end{aligned}$$

where a_i denotes the chemical concentration of species i in the solution, b_i the chemical concentration of species i in the rock, λ_i the reaction rate for species i to react with the rock, μ_i the reaction rate for the rock to release species i, V the advection speed, ϵ the diffusion coefficient and δ and γ denote reaction constants. We assume that $\epsilon = 0$ in our investigation of this model. The conserved quantities are

$$d := a_1 + b_1 + a_3 + b_3, \qquad e := a_1 + b_1 - a_2 - b_2, \tag{5.7}$$

and they satisfy

$$d_t + V(a_1 + a_3)_x = 0, \qquad e_t + V(a_1 - a_2)_x = 0. \tag{5.8}$$

In (Budd et al. 1997) a numerical two-step method was considered for this system. Firstly, a transport step was made which ignored the chemical reactions; then local chemical equilibrium was assumed. This allowed the reactions to be decoupled from the transport equations so that an operator splitting technique could be applied. In our terminology this procedure leads to the Equilibrium model.

As discussed in the Introduction, a particular case of interest (as is considered in (Budd et al. 1997)) is when one or more species are flushed through the system. This means there is not much retardation present. We suppose that a_3 is flushed through, but a similar model could be obtained for the flushing through of either a_1 or a_2. Since a_3 is flushed through the system there is no equation for b_3 (a_3 is not adsorbed into the rock at any stage but is simply carried along by the groundwater). We also make a further simplification by setting $a_1 = a_2$ and $b_1 = b_2$. Then (5.1)–(5.6) reduce to a three component model for a_1, b_1 and a_3. Re-labelling these as a, b and c, respectively, we obtain the *Flushing-through Model* as defined in (1.17)–(1.19) in Chapter 1, which we again state here for convenience

$$a_t + Va_x = -\lambda a + \mu b - \gamma a^2 + \delta c \tag{5.9}$$

$$b_t = \lambda a - \mu b \tag{5.10}$$

$$c_t + Vc_x = \gamma a^2 - \delta c. \tag{5.11}$$

Observe that (5.10) only involves a and b and (5.11) only involves a and c. The conserved quantity d is now $d := a + b + c$ and satisfies

$$d_t + V(a + c)_x = 0. \tag{5.12}$$

We need to specify initial and boundary conditions for this model which we will take as standard throughout this Chapter. As for the Linear Model, the concentrations are assumed to be zero along the $t = 0$ axis. Once the boundary condition for a is specified, the boundary condition for b follows directly. We could think of the concentration c as being zero both on the $x = 0$ and $t = 0$ boundaries and then being formed directly from a using (5.11). However, considering the situation when the system is in equilibrium, the right hand side of both (5.10) and (5.11) will be set to zero. Hence $c(0, t)$ is assumed to satisfy this equilibrium condition and so

$$a(x, 0) = b(x, 0) = c(x, 0) = 0, \tag{5.13}$$

with

$$a(0, t) = g(t), \quad b(0, t) = \lambda \int_0^t g(r) e^{-\mu(t-r)} \, \mathrm{d}r, \quad c(0, t) = \frac{\gamma}{\delta} g(t)^2. \tag{5.14}$$

We now apply some of the analytical techniques from Chapters 2 and 3 to give a better understanding of this system. Firstly, both the Equilibrium and Improved-equilibrium models are derived, using the method described in (Chen et al. 1994). This again shows the diffusion present in this model but, as for the Linear Model, the resulting reduced system does not accurately describe the physical aspects except in the extreme situation when all the parameters are large and of the same order. There only needs to be one of a different magnitude for the Improved-equilibrium model to be inadequate; numerical results will illustrate this.

A key feature of the Flushing-through model is the fact that it is possible for the concentrations a and c to move at different speeds. This cannot be deduced from either the Equilibrium or Improved-equilibrium models. However, in Section 5.4 we will perform a modified equation analysis of the weighted box-trap scheme applied to (5.9)–(5.11) which can be used to show this phenomenon (and is valid for all sizes of the parameters). The key equations are given in (5.60) and (5.61); these involve a pair of partial differential equations for the numerical approximations A and C, provided λ and μ are assumed large. These will immediately show why the concentrations a and c could move at different speeds and can be used to deduce other interesting features observed from numerical experiments carried out in Section 5.3.

5.2 The Improved-equilibrium model

In Chapter 2 we derived a single convection diffusion equation which approximates a system of hyperbolic conservation laws with relaxation. This was then applied to the Linear Model. We now follow the same procedure for The Flushing-through Model. Equations (5.9), (5.10) and (5.11) can be written as

$$\mathbf{u}_t + A\mathbf{u}_x + \frac{1}{\epsilon}\mathbf{S}(\mathbf{u}) = 0, \tag{5.15}$$

with

$$\mathbf{u} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \quad A = \begin{bmatrix} V & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & V \end{bmatrix}, \quad \mathbf{S}(\mathbf{u}) = \begin{bmatrix} a - \frac{\mu}{\lambda}b + \frac{\gamma}{\lambda}a^2 - \frac{\delta}{\lambda}c \\ -a + \frac{\mu}{\lambda}b \\ -\frac{\gamma}{\lambda}a^2 + \frac{\delta}{\lambda}c \end{bmatrix}, \tag{5.16}$$

where $\epsilon := 1/\lambda$, with $\lambda >> 1$. Let $Q = [1, 1, 1]$ and then the condition (2.39) holds, i.e. $Q\mathbf{S}(\mathbf{u}) = \mathbf{0}$. Note that we make a change of notation from that in Section 2.5 of Chapter 2: the set ω, defined in (2.42), becomes

$$\omega := \{d \in \mathbb{R} \mid d = Q\mathbf{u}\}, \tag{5.17}$$

and so d replaces c in the previous terminology. If we set

$$\mathcal{E}(d) = \left[\alpha_1(d),\ \alpha_2(d),\ \alpha_3(d)\right]^T, \tag{5.18}$$

then from (2.44), which says $Q\mathcal{E}(d) = d$

$$\alpha_1(d) + \alpha_2(d) + \alpha_3(d) = d. \tag{5.19}$$

We also need (2.43) to hold (i.e. $\mathbf{S}(\mathcal{E}(d)) = 0$) and so

$$\begin{aligned} \alpha_1(d) - \frac{\mu}{\lambda}\alpha_2(d) + \frac{\gamma}{\lambda}\alpha_1(d)^2 - \frac{\delta}{\lambda}\alpha_3(d) &= 0 && (5.20)\\ -\alpha_1(d) + \frac{\mu}{\lambda}\alpha_2(d) &= 0 && (5.21)\\ -\frac{\gamma}{\lambda}\alpha_1(d)^2 + \frac{\delta}{\lambda}\alpha_3(d) &= 0. && (5.22) \end{aligned}$$

Equations (5.19), (5.20)–(5.22) can be solved to find α_1, α_2 and α_3. We obtain a quadratic to solve for α_1 given by

$$\mu K\alpha_1(d)^2 + (\lambda + \mu)\alpha_1(d) - \mu d = 0, \tag{5.23}$$

where

$$K := \frac{\gamma}{\delta}. \tag{5.24}$$

The flux function g (from (2.47)) for the Equilibrium model (2.46) is simply

$$g(d) = QA\mathcal{E}(d) = V\left(\alpha_1(d) + \alpha_3(d)\right). \tag{5.25}$$

We can solve (5.23) to find α_1 (taking the positive square root) and then use (5.22) to find α_3. Hence we obtain the Equilibrium model

$$d_t + V(g(d))_x = 0, \tag{5.26}$$

where

$$g(d) = d + \frac{\lambda}{2\mu^2 K}\left((\lambda + \mu) - \sqrt{(\lambda + \mu)^2 + 4\mu^2 K d}\right). \tag{5.27}$$

This is a nonlinear conservation law and so, depending on the initial and boundary data, a shock could form. We could solve (5.26) numerically by applying the box scheme; the corrected algorithm (as described in Chapter 3) could be used when the data is either a shock or shock-forming. Then a, b and c can easily be found by using the equilibrium manifold $\mathcal{M} := \{u \mid R(u) = 0\}$ and the fact that $d = a + b + c$.

However, as for the Linear Model, this is not a very good approximation to the Flushing-through model unless all the parameters are large (which numerical results will illustrate

in Section 5.3.3). A correction to (5.26) can be made; equations (2.52) and (2.54) are used to find $\mathcal{M}^{(1)}[d]$. Since $\mathbf{f}(\mathbf{u}) = A\mathbf{u}$ in this case, with A from (5.16), we have

$$\left[I - (\mathcal{M}[d])_d Q\right](f(\mathcal{E}(d)))_x = \begin{bmatrix} V\alpha_1'(d)\left[1 - \alpha_1'(d) - \alpha_3'(d)\right] d_x \\ -V\alpha_2'(d)\left[\alpha_1'(d) + \alpha_3'(d)\right] d_x \\ V\alpha_3'(d)\left[1 - \alpha_1'(d) - \alpha_3'(d)\right] d_x \end{bmatrix}$$

where we have used the fact that $\alpha_1'(d)+\alpha_2'(d)+\alpha_3'(d) = 1$ (obtained from differentiating (5.19) with respect to d). Also

$$\frac{\partial}{\partial \mathbf{u}}(\mathbf{S}(\mathcal{E}(d))) = \begin{bmatrix} 1 + \frac{2\gamma}{\lambda}\alpha_1(d) & -\frac{\mu}{\lambda} & -\frac{\delta}{\lambda} \\ -1 & \frac{\mu}{\lambda} & 0 \\ -\frac{2\gamma}{\lambda}\alpha_1(d) & 0 & \frac{\delta}{\lambda} \end{bmatrix}.$$

Following (Liu 1987) we set

$$\mathcal{M}^{(1)}[d] = [\beta_1(d),\, \beta_2(d),\, \beta_3(d)]^T, \tag{5.28}$$

where β_1, β_2 and β_3 need to be found by solving the system (2.52), i.e.

$$\frac{\partial}{\partial \mathbf{u}}(\mathbf{S}(\mathcal{E}(d)))\mathcal{M}^{(1)}[d] = \left[I - (\mathcal{M}[d])_d Q\right](f(\mathcal{E}(d)))_x, \tag{5.29}$$

and using the condition (2.54) ($Q\mathcal{M}^{(1)}[d] = 0$), which is now

$$\beta_1(d) + \beta_2(d) + \beta_3(d) = 0.$$

The details are not given here but, after some manipulation, the first-order correction to the local equilibrium approximation (c.f. (2.55)) is given by

$$d_t + \left(QA\mathcal{M}^{\epsilon}[d]\right)_x = 0,$$

where

$$QA\mathcal{M}^{\epsilon}[d] = QA\left(\mathcal{E}(d) + \epsilon\mathcal{M}^{(1)}[d]\right) = V\left[\alpha_1(d) + \alpha_3(d)\right] + \epsilon V\left[\beta_1(d) + \beta_3(d)\right].$$

Hence the Improved-equilibrium model is

$$d_t + V\left[\alpha_1(d) + \alpha_3(d)\right]_x + \epsilon V\left[\beta_1(d) + \beta_3(d)\right]_x = 0. \tag{5.30}$$

This is finally written as

$$d_t + g(d)_x - \left(\phi(d)d_x\right)_x = 0, \tag{5.31}$$

where $g(d)$ is defined in (5.27) and $\phi(d)$ is given by

$$\phi(d) = \frac{V^2\lambda/\delta\mu}{[(\lambda+\mu)^2+4\mu^2Kd]}\left(\lambda(\mu-2\delta)+\frac{\lambda[\delta\lambda-\mu(\lambda+\mu)]}{\sqrt{(\lambda+\mu)^2+4\mu^2Kd}}+\delta\sqrt{(\lambda+\mu)^2+4\mu^2Kd}\right). \tag{5.32}$$

Unfortunately, this equation is highly nonlinear and therefore not easy to analyse. In Section 5.3.3 we numerically compare the Equilibrium model, the Improved-equilibrium model and the Flushing-through Model for various sizes of the parameters. We will show that (5.31) is only a reasonable approximation if all the parameters (i.e. λ, μ, γ and δ) are large, which is consistent with the theory, but is very restrictive in practice. It does, however, highlight the diffusion present in the system.

5.2.1 Alternative form of the Equilibrium model

In the previous Section we derived the Equilibrium and Improved-equilibrium models in terms of the conserved quantity d. However, the Equilibrium model could have been derived in terms of a. Suppose the reaction rates are set to be infinite. Then

$$0 = \lambda a - \mu b, \qquad 0 = c - Ka^2. \tag{5.33}$$

These can be used to obtain an expression for d in terms of a, and so

$$d = \frac{\lambda+\mu}{\mu}a + Ka^2. \tag{5.34}$$

Now substituting (5.34) into the conservation equation (5.12) gives the Equilibrium model entirely in terms of a, i.e.

$$\left(\frac{\lambda+\mu}{\mu}a+Ka^2\right)_t + V\left(a+Ka^2\right)_x = 0. \tag{5.35}$$

This equation can be written as

$$\eta(a)_t + \psi(a)_x = 0, \tag{5.36}$$

where

$$\eta(a) := \frac{\lambda+\mu}{\mu}a + Ka^2, \qquad \psi(a) := V\left(a+Ka^2\right). \tag{5.37}$$

As discussed in (LeVeque 1992, page 37), since ψ is convex, $\eta(a)$ and $\psi(a)$ are the entropy function and entropy flux respectively. Provided a is smooth (and $\eta'(a) \neq 0$),

the conservation law (5.36) can be written as

$$a_t + \frac{\psi'(a)}{\eta'(a)} a_x = 0,$$

where

$$\frac{\psi'(a)}{\eta'(a)} = V\left(\frac{\mu + 2\mu K a}{\lambda + \mu + 2\mu K a}\right) =: f'(a). \tag{5.38}$$

Then (5.35) becomes

$$a_t + V\left(\frac{\mu + 2\mu K a}{\lambda + \mu + 2\mu K a}\right) a_x = 0. \tag{5.39}$$

We could linearise $f'(a)$ about a fixed point to make this equation easier to study. This idea will be examined in more detail in Section 5.4; linearising the modified equation expansion (which can be reduced to a similar equation when λ and μ are large) will enable us to make deductions about the different speeds that can arise in the Flushing-through Model.

The equation (5.39) can also be written in conservative form as

$$a_t + (f(a))_x = 0, \tag{5.40}$$

where $f(a)$ can be found by integrating (5.38). A simple calculation gives

$$f(a) = V\left(a - \frac{\lambda}{2\mu K} \ln\left[\lambda + \mu + 2\mu K a\right]\right). \tag{5.41}$$

This is an alternative form of the Equilibrium model, in terms of a rather than d (from (5.26)). These forms cannot be linked in the same way as for the Linear Model: in that case the Equilibrium models for a and the conserved quantity c actually satisfied the same equation. However, the analogous equations (5.26) and (5.39) have different weak solutions. We need to know which quantity is being conserved otherwise the resulting equation will not be valid in the limit as the parameters tend to their equilibrium values. Although not advisable (except in extreme situations), if the Equilibrium model is used as an approximation to a general system, it should always be in terms of the conserved quantity. Also, to enable comparisons between the Equilibrium and Improved-equilibrium models we have to use the conserved quantity as the latter can only be obtained in this form. This will be done for the Flushing-through Model in Section 5.3.3.

5.3 The weighted box-trap scheme

The Flushing-through Model is now written as the conservation equation (5.12) coupled with the source equations (5.10) and (5.11), to be consistent with the formation of the

box-trap scheme for the Linear Model in Chapter 3. The weighted box scheme is then applied to (5.12) and (5.11) and the trapezoidal scheme to (5.10).

There is now a source term in equation (5.11) of the model. Cunge & Holly Jr (1980, pages 92-93) discuss the discretisation of nonlinear terms and coefficients using the box scheme. If we wish to apply the box scheme to the right hand side of (5.11), we could follow (Cunge & Holly Jr 1980) and so the discretisation would be

$$\mu_t\mu_x(\gamma A^2 - \delta C). \tag{5.42}$$

However, we now wish to apply the weighted box scheme to (5.11); it is not immediately clear whether to alter (5.42) by changing μ_t to θ_t, or simply to use θ_t for the time averaging in the spatial derivative. Cunge & Holly Jr (1980, pages 92-93) use θ_t for their source terms, although what they refer to as the *Preissman box scheme* always uses a θ_t weighting. If we were interested in the steady state problem for (5.11) then it would be important to time average the c_x term and the source term in the same way. However, the Flushing-through Model is an extension of the Linear Model where the weighted box-trap scheme involves discretising the reaction equation using the trapezoidal rule. To be consistent with this form (and to ensure consistency between both the source terms in (5.10) and (5.11) in the Flushing-through Model) we will use μ_t for time average of the source term. This is not unreasonable since using $\theta > \frac{1}{2}$ introduces extra diffusion into the numerical solution, which we wish to avoid.

Hence, the discretisations of (5.12), (5.10) and (5.11) are given by

$$\begin{aligned}(1+2\theta p)A_{j+1}^{n+1} + (1-2\theta p)A_j^{n+1} - [1-2(1-\theta)p]A_{j+1}^n - [1+2(1-\theta)p]A_j^n \\ +(1+2\theta p)C_{j+1}^{n+1} + (1-2\theta p)C_j^{n+1} - [1-2(1-\theta)p]C_{j+1}^n - [1+2(1-\theta)p]C_j^n \\ +(B_{j+1}^{n+1} - B_{j+1}^n) + (B_j^{n+1} - B_j^n) = 0,\end{aligned} \tag{5.43}$$

$$B_{j+1}^{n+1} = \frac{\frac{1}{2}\lambda'}{1+\frac{1}{2}\mu'}(A_{j+1}^{n+1} + A_{j+1}^n) + \left(\frac{1-\frac{1}{2}\mu'}{1+\frac{1}{2}\mu'}\right) B_{j+1}^n, \tag{5.44}$$

and

$$\begin{aligned}\left(1+2\theta p+\tfrac{1}{2}\delta'\right)C_{j+1}^{n+1} + \left(1-2\theta p+\tfrac{1}{2}\delta'\right)C_j^{n+1} \\ -\left[1-2(1-\theta)p-\tfrac{1}{2}\delta'\right]C_{j+1}^n - \left[1+2(1-\theta)p-\tfrac{1}{2}\delta'\right)\Big]C_j^n \\ = \tfrac{1}{2}\gamma'\left[{A_{j+1}^{n+1}}^2 + {A_j^{n+1}}^2 + {A_{j+1}^n}^2 + {A_j^n}^2\right].\end{aligned} \tag{5.45}$$

Using (5.44) and (5.45) to eliminate B_{j+1}^{n+1} and C_{j+1}^{n+1} from (5.43) leads to a quadratic to solve for A_{j+1}^{n+1} (where we take the positive solution). This value can then be substituted into (5.44) and (5.45) to find B_{j+1}^{n+1} and C_{j+1}^{n+1} respectively.

5.3.1 Numerical results

For simplicity, suppose the boundary condition for a is the smooth Gaussian pulse

$$g(t) = e^{-25(t-0.5)^2}. \tag{5.46}$$

Then we can assume $\theta = \frac{1}{2}$. Budd et al. (1997) test their numerical solution methods for K ($:= \gamma/\delta$) small, of order 1, and large; we also consider a wide range of values for γ and δ as well as λ and μ (but restricting $\lambda \geq \mu$). These situations are all physically realistic. Since c is not present in the rock in the Flushing-through Model there will not be much retardation for this concentration and so we expect c to move faster than a. On examining equations (5.9)–(5.11) we see that a is fed by the boundary data and the $\mu b + \delta c$ term and damped by the $\lambda a + \gamma a^2$ term; also, c is fed by the boundary data and the γa^2 term and damped by the δc term.

Firstly, consider the situation when the source terms are not dominant. This is shown in Figure 5-1 where the concentrations a, b and c are plotted for four cases of λ, μ, γ and δ small. As expected, in all these experiments the solutions move at the advected speed. From the top left plots we observe that a is damped more than c (because $\lambda a + \gamma a^2$ is larger than δc). If λ is increased to 3 (top right plots) then both a and c are reduced and b is increased; c is fed by a but this has been damped further. In the bottom sets of plots, λ and μ are fixed at 1 and γ and δ are varied. When $\delta < \gamma$ (bottom left) a and c are not as damped since γa^2 has increased and this feeds c. When $\delta > \gamma$ (bottom right) the situation is similar but c has become smaller since there is more damping from the δc term.

In Figure 5-2 we fix λ and μ large and vary γ and δ. Since there is significantly more damping in these plots, the boundary conditions are now omitted. In the top two sets of plots γ and δ are still small and we observe some of our earlier predictions: for $\gamma > \delta$ (top left), a moves at the reduced speed V' and c still moves at speed V. The concentrations a and b are both zero until c is switched on. However, when $\delta > \gamma$ (top right) we observe that c has two peaks, one of which is moving with speed V and the other is at a much slower speed (but not quite at the reduced speed). However, note that c is much smaller in the bottom right plot than in the bottom left, which is to be expected since the δc term is large.

Finally, the bottom sets of plots in Figure 5-2 show both γ and δ large. When $\gamma > \delta$ (bottom left) some retardation is now observed for c: it is moving at a slower speed than V, though is still faster than a. It is interesting to note that in the left two sets of plots in Figure 5-2 the value of K is the same but the results are very different; a is travelling at a much larger speed in the bottom set which is not easy to deduce from the equations. We would expect a to travel at the reduced speed so this change must

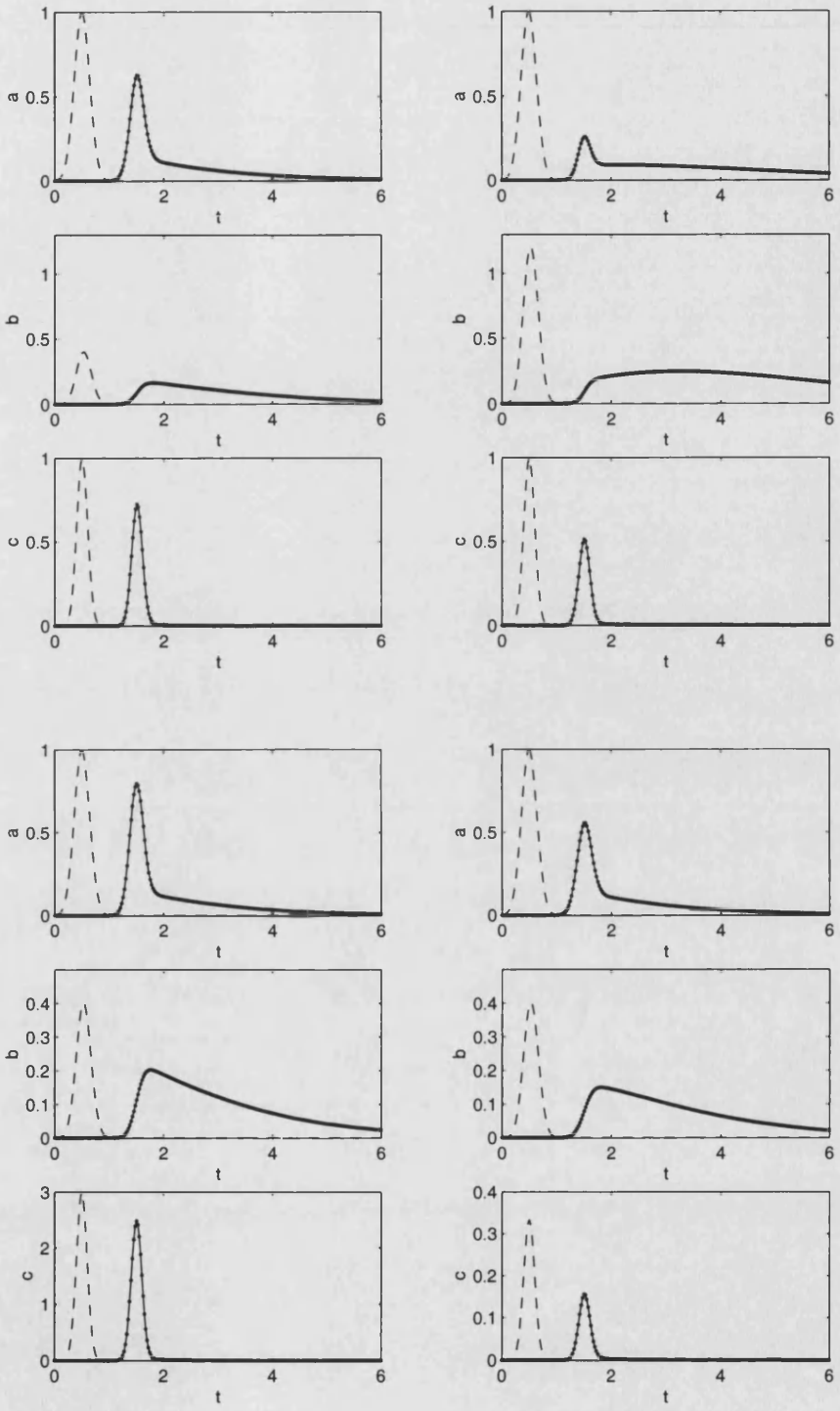

Figure 5-1: *The box-trap scheme (dots joined by an unbroken line) applied to the Flushing-through Model plotted against t at fixed $x = 1$ (with $V = 1$). Four cases each showing a, b and c; parameters are $\lambda = \mu = \gamma = \delta = 1$ (top left) with one change in other cases: $\lambda = 3$ (top right), $\gamma = 3$ (bottom left) and $\delta = 3$ (bottom right). The boundary data is a Gaussian curve (shown as a dashed line)*

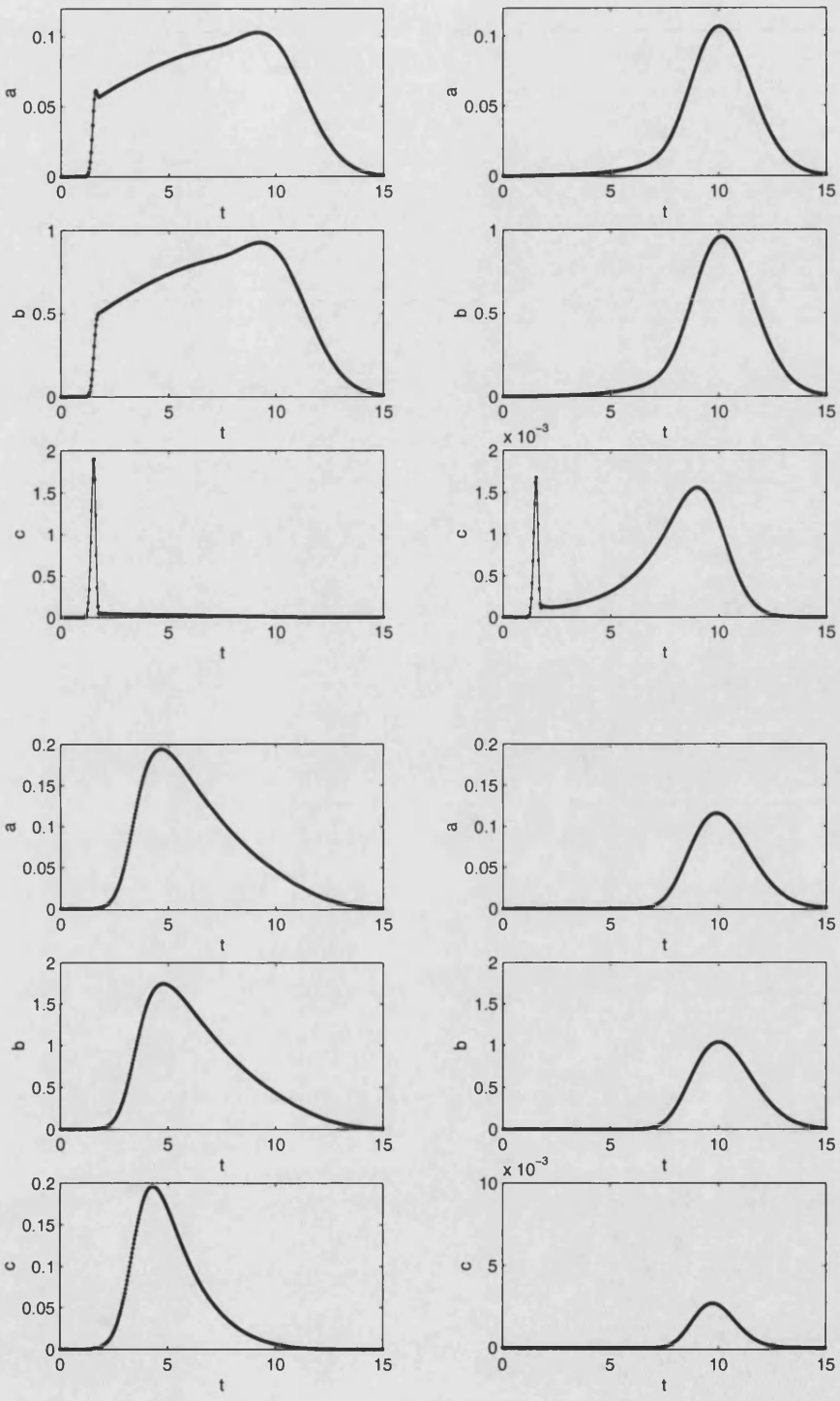

Figure 5-2: *The box-trap scheme (dots joined by an unbroken line) applied to the Flushing-through Model plotted against t at fixed $x = 1$ (with $V = 1$). Four cases each showing a, b and c with $\lambda = 90$ and $\mu = 10$. Other parameters are $\gamma = 5$, $\delta = 1$ (top left), $\gamma = 1$, $\delta = 5$ (top right), $\gamma = 50$, $\delta = 10$ (bottom left) and $\gamma = 10$, $\delta = 50$ (bottom right).*

be due to the δc term becoming dominant. The bottom right plots show $\gamma < \delta$ and all the concentrations again move at the reduced speed. The concentration c is now very heavily damped by the δc term. We know that it is also fed by γa^2 but γ is not as large and so does not have as much effect as the damping term.

In conclusion, when λ and μ are fixed to be large, a and b move at the reduced speed except when γ and δ (with $\gamma > \delta$) are also large (see the bottom left plots in Figure 5-2). We will explain this using a modified equation analysis in Section 5.4. Also, c moves at almost the reduced speed for all sizes of $\gamma < \delta$ (though there is an extra peak at the advected speed when these parameters are small).

5.3.2 Varying θ and the CFL number for non-smooth data

We now suppose that $g(t)$ is a square pulse and discuss how the weighted box-trap scheme copes when there are varying sizes of the parameters. We would like to choose the mesh to match the correct speed (so the solution travels along the numerical characteristic). This was easy in the Linear Model as there was only one speed to consider for a given value of λ and μ: if λ and μ are small, choose $p = 1$ (where p is the CFL number) and if they are large choose $p' := \frac{\mu}{\lambda+\mu} p = 1$. However, for the Flushing-through Model this is not so simple. We want the weighted box-trap scheme to be robust enough to cope with these large parameter ranges. Results are now presented for a large range of sizes of the parameters and we will aim to predict in which situations we need to take a small Δt (corresponding to $p = 1$) rather than p' closer to 1. We will also investigate when we need to use $\theta > \frac{1}{2}$ to reduce any oscillations.

In Figure 5-3 we consider $\lambda = \mu = \gamma = \delta = 1$. Since all the parameters are small we should take $p = 1$. In the left plots $\theta = \frac{1}{2}$ and we observe oscillations for a and c. These do not propagate because, for the chosen value of p, the dispersion term is zero (as discussed for the linear advection equation and the Linear Model in Chapter 3). If θ is increased slightly (right plots) the oscillations have disappeared. As predicted, when all the parameters are small, setting $p = 1$ and using $\theta > \frac{1}{2}$ gives the best results.

Figure 5-4 shows results for $\lambda = 90$, $\mu = 10$, $\gamma = \delta = 1$ and $\theta = \frac{1}{2}$. The concentrations a and b are moving at the reduced speed but c is moving at speed V. In the left plots $p = 1$ and in the right $p' = 1$. The solution for c is highly oscillatory and so we have to use $p = 1$. There are oscillations for a and c but these can be eliminated by increasing θ from $\frac{1}{2}$. Figures 5-5 and 5-6 show the same λ and μ but γ and δ are increased by a small amount. When $\gamma > \delta$ (Figure 5-5) we must take $p = 1$ but, if $\gamma < \delta$ (Figure 5-6) we can take a larger time step and only need $p' = 1$ to eliminate the oscillations (and can use $\theta = \frac{1}{2}$). This is because c is moving closer to the reduced speed.

It is interesting to compare Figure 5-5 with the top left plots in Figure 5-2; they have the

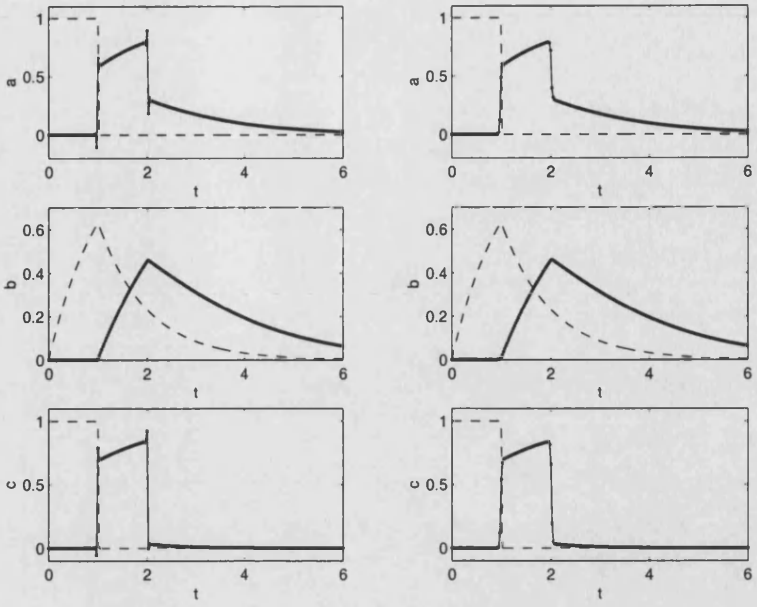

Figure 5-3: *The box-trap scheme (dots joined by an unbroken line) applied to the Flushing-through Model plotted against t at fixed $x = 1$, with $\lambda = \mu = \gamma = \delta = 1$ and $p = 1$. In the left plots $\theta = 0.5$ and in the right plots $\theta = 0.51$. The boundary data is shown as a dashed line. The plots show a (top), b (middle) and c (bottom).*

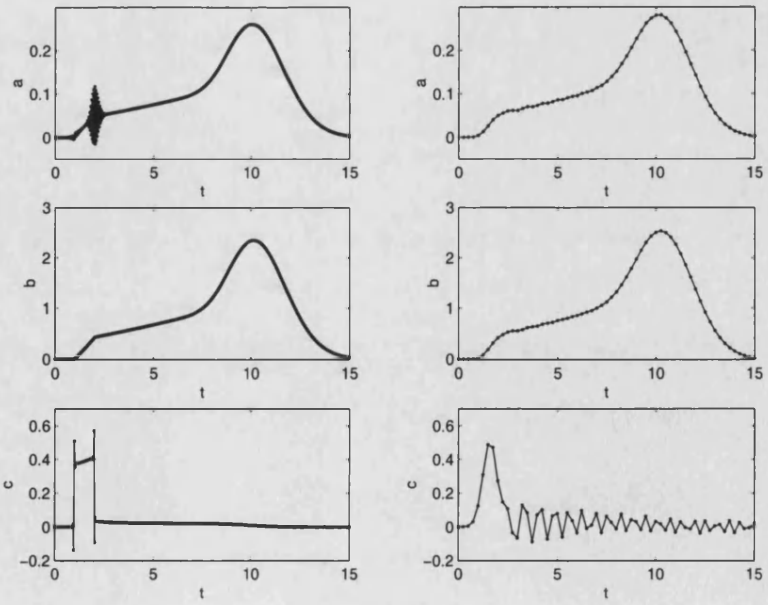

Figure 5-4: *The box-trap scheme (dots joined by an unbroken line) applied to the Flushing-through Model plotted against t at fixed $x = 1$, with $\lambda = 90$, $\mu = 10$, $\gamma = \delta = 1$ and $\theta = 0.5$. In the left plots $p = 1$ and in the right plots $p' = 1$.*

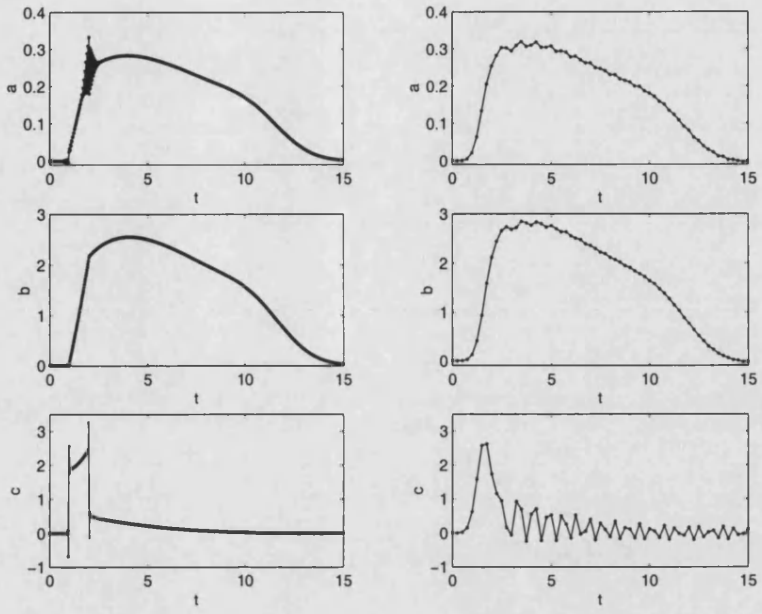

Figure 5-5: *The box-trap scheme (dots joined by an unbroken line) applied to the Flushing-through Model plotted against t at fixed $x = 1$, with $\lambda = 90$, $\mu = 10$, $\gamma = 5$, $\delta = 1$ and $\theta = 0.5$. In the left plots $p = 1$ and in the right plots $p' = 1$.*

same values for the parameters but different boundary conditions. When the boundary data is smooth the peak is close to the reduced speed but when a square pulse is used the speed at which a and b move is much faster. If γ and δ are increased (with K still fixed at 5) the peaks all move at a faster speed than V' whichever boundary data is used (shown for smooth data in the top left plots in Figure 5-2). This phenomena will again be discussed in Section 5.4. When γ and δ are increased further we can always take $p' = 1$ and there are no oscillations. Also, if λ and μ are small with γ and δ much larger we always need to take $p = 1$ and increase θ from $\frac{1}{2}$, as shown in Figure 5-7.

In conclusion, the weighted box-trap scheme is robust enough to cope with varying sizes of the parameters (and therefore varying speeds in the system). The results in Figure 5-3 indicate that when all of the parameters are small we generally have to choose the mesh which matches the fastest characteristic (i.e. $p = 1$) and choose $\theta > \frac{1}{2}$ (but as close to $\frac{1}{2}$ as possible to prevent too much diffusion caused by the weighting). If some of the parameters are increased (but not all), again we need to take $p = 1$ and $\theta > \frac{1}{2}$; although there is an exception when $\gamma < \delta$, as we saw in Figure 5-6. As soon as all the parameters become large we are able to choose the mesh to match direction of the slowest speed and do not need to use any weighting.

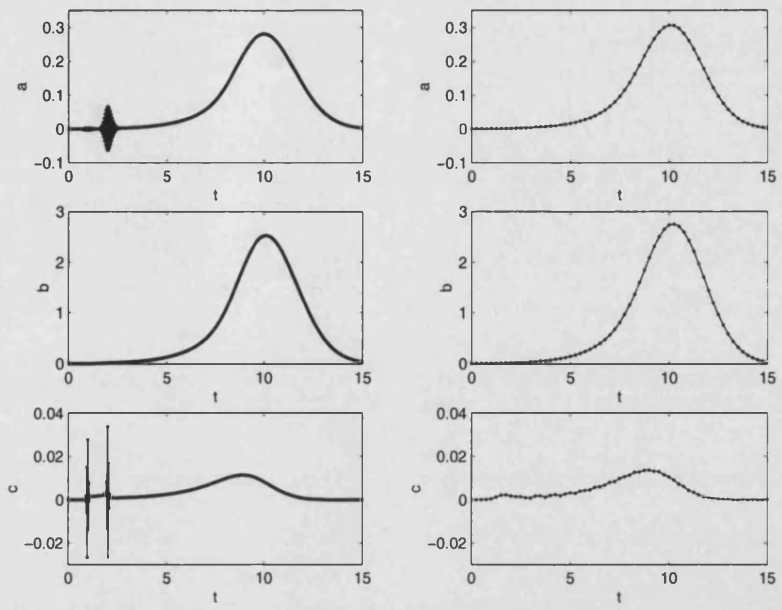

Figure 5-6: *The box-trap scheme (dots joined by an unbroken line) applied to the Flushing-through Model plotted against t at fixed $x = 1$, with $\lambda = 90$, $\mu = 10$, $\gamma = 1$, $\delta = 5$ and $\theta = 0.5$. In the left plots $p = 1$ and in the right plots $p' = 1$.*

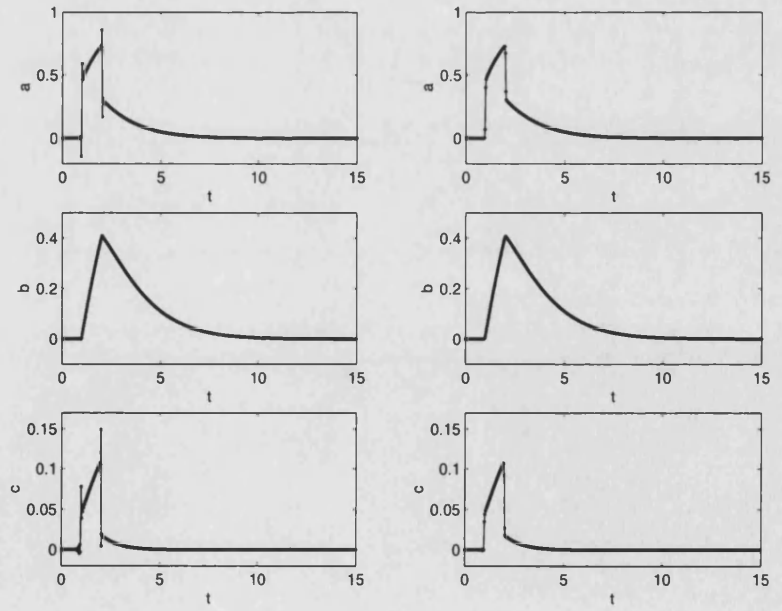

Figure 5-7: *The box-trap scheme (dots joined by an unbroken line) applied to the Flushing-through Model plotted against t at fixed $x = 1$, with $\lambda = 1$, $\mu = 1$, $\gamma = 10$, $\delta = 50$ and $p = 1$. In the left plots $\theta = 0.5$ and in the right plots $\theta = 0.51$.*

5.3.3 Comparison of the three models

Suppose the Equilibrium Model (5.26) and Improved-equilibrium model (5.31) are approximated using simple explicit schemes: the former by the upwind scheme, and so

$$d_j^{n+1} = d_j^n - \nu\big(g(d_j^n) - g(d_{j-1}^n)\big), \tag{5.47}$$

and the latter by the central scheme

$$d_j^{n+1} = d_j^n - \tfrac{1}{2}\nu\big(g(d_{j+1}^n) - g(d_{j-1}^n)\big) + \xi\left[(d_{j+1}^n - d_j^n)\phi(d_{j+1/2}^n) - (d_j^n - d_{j-1}^n)\phi(d_{j-1/2}^n)\right], \tag{5.48}$$

where

$$d_{j-1/2}^n = \tfrac{1}{2}(d_{j-1}^n + d_j^n), \qquad d_{j+1/2}^n = \tfrac{1}{2}(d_{j+1}^n + d_j^n).$$

Note that ν is the mesh ratio and $\xi = \triangle t/\triangle x^2$. Neither of these numerical schemes give very good approximations unless the mesh ratio is very small; they introduce a lot of diffusion into the numerical solution. However, they are adequate for our purposes as we simply wish to compare these models with the weighted box-trap scheme applied to the Flushing-through Model qualitatively, and not the numerical accuracy.

Figures 5-8 and 5-9 show comparisons of these three models for various parameters. The top left plot in Figure 5-9 illustrates that when all the parameters are small the Improved-equilibrium model gives a very bad approximation and is severely damped (as we would expect). The top right shows that increasing λ and μ alone does not give much improvement. In Figure 5-8 we have fixed $\lambda = 90$ and $\mu = 10$ and varied γ and δ. The Improved-equilibrium model does give a reasonable approximation when $\gamma < \delta$, which improves as they increase but K kept fixed (see the bottom right plot). However, it is only when μ is increased from 10 that the Improved-equilibrium model becomes more accurate and is noticeably better than the Equilibrium model. This can be seen in the bottom two plots in Figure 5-9.

We conclude that the Equilibrium and Improved-equilibrium models only give reasonable approximations to the Flushing-through Model when all the parameters are large (and of the same order), which is shown in the bottom plots in Figure 5-9. This is to be expected since we needed $\epsilon := 1/\lambda$ to be large in order to ignore higher order terms in the derivation. However, there is an exception in the bottom right plot in Figure 5-8: the Improved-equilibrium model is quite accurate even though there is quite a difference between both λ, μ and γ, δ.

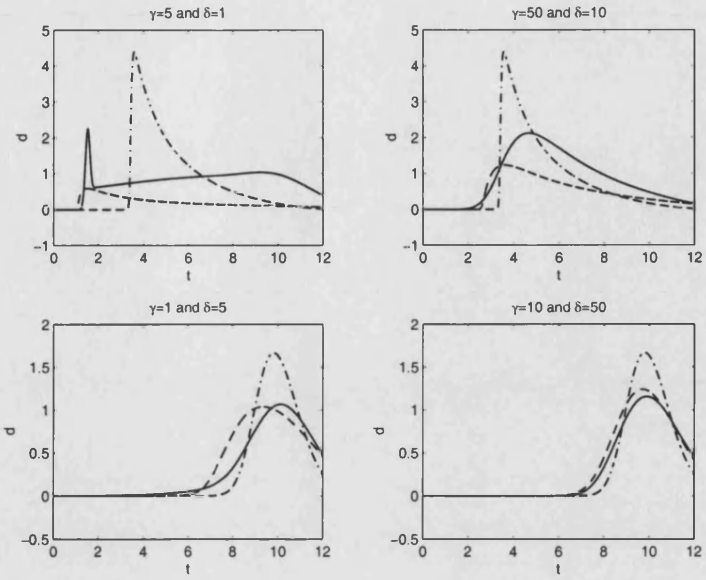

Figure 5-8: *The Equilibrium model (dot-dashed line), the Improved-equilibrium model (dashed line) and the box-trap scheme applied to the Flushing-through Model (the solid line). In all four plots* $\lambda = 90$ *and* $\mu = 10$ *and in the top left* $\gamma = 5$ *and* $\delta = 1$*, top right* $\gamma = 50$ *and* $\delta = 10$*, bottom left* $\gamma = 1$ *and* $\delta = 5$ *and bottom right* $\gamma = 10$ *and* $\delta = 50$.

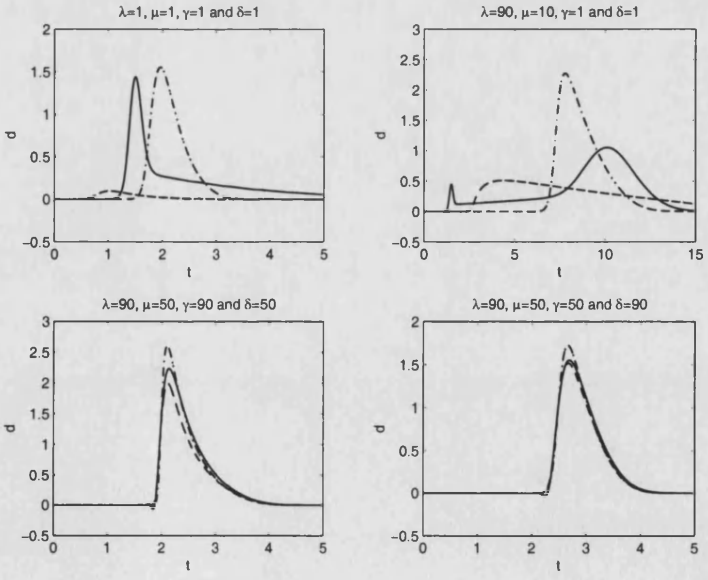

Figure 5-9: *The Equilibrium model (dot-dashed line), the Improved-equilibrium model (dashed line) and the box-trap scheme applied to the Flushing-through Model (the solid line). In the top left plot* $\lambda = \mu = \gamma = \delta = 1$*, top right* $\lambda = 90$*,* $\mu = 10$ *and* $\gamma = \delta = 1$*, bottom left* $\lambda = 90$*,* $\mu = 50$*,* $\gamma = 90$ *and* $\delta = 50$ *and bottom right* $\lambda = 90$*,* $\mu = 50$*,* $\gamma = 50$ *and* $\delta = 90$.

5.4 Modified equation analysis

Consider the discretised equations written in terms of finite difference operators

$$\mu_x\delta_t A + p\theta_t\delta_x A = \mu_x\mu_t\left(-\lambda' A + \mu' B + \delta' C\right) - \gamma'\mu_x\mu_t(A^2) \quad (5.49)$$

$$\delta_t B = \mu_t\left(\lambda' A - \mu' B\right) \quad (5.50)$$

$$\mu_x\delta_t C + p\theta_t\delta_x C = -\delta'\mu_x\mu_t C + \gamma'\mu_x\mu_t(A^2). \quad (5.51)$$

These can be written in terms of the operators $\mathcal{D}_t$ and $\mathcal{D}_x$ defined in (3.90), i.e.

$$\mathcal{D}_t := \frac{\delta_t\mu_t^{-1}}{\Delta t}, \qquad \mathcal{D}_x := \frac{\delta_x\mu_x^{-1}}{\Delta x}.$$

Then (5.49), (5.50) and (5.51) become

$$\mathcal{D}_t A + V\mathcal{M}_t\mathcal{D}_x A = -\lambda A + \mu B + \delta C - \gamma A^2 \quad (5.52)$$

$$\mathcal{D}_t B = \lambda A - \mu B \quad (5.53)$$

$$\mathcal{D}_t C + V\mathcal{M}_t\mathcal{D}_x C = -\delta C + \gamma A^2, \quad (5.54)$$

where $\mathcal{M}_t := \theta_t^{-1}\mu_t$. We have omitted $(\tilde{A})$ for the prolongation of the mesh function A (and similarly for B and C) to simplify the notation. Now define

$$L_1 := \mathcal{D}_t + V\mathcal{M}_t\mathcal{D}_x + \lambda, \qquad L_2 := \mathcal{D}_t + \mu, \qquad L_3 := \mathcal{D}_t + V\mathcal{M}_t\mathcal{D}_x + \delta. \quad (5.55)$$

Hence (5.52)–(5.54) can be written as

$$L_1 A - \mu B + \gamma A^2 - \delta C = 0 \quad (5.56)$$

$$L_2 B = \lambda A \quad (5.57)$$

$$L_3 C = \gamma A^2. \quad (5.58)$$

Multiplying (5.56) by L_2 and substituting (5.57) into the resulting equation leads to a relation in terms of A and C only

$$\left[\frac{\mathcal{D}_t^2}{\lambda+\mu} + \left(1 + \frac{V\mathcal{M}_t\mathcal{D}_x}{\lambda+\mu}\right)\mathcal{D}_t + \frac{V\mu\mathcal{D}_x}{\lambda+\mu}\right] A + \left(\frac{\mathcal{D}_t+\mu}{\lambda+\mu}\right)\left(\gamma A^2 - \delta C\right) = 0. \quad (5.59)$$

When both λ and μ are large it is convenient to neglect $O(1/(\lambda+\mu))$ terms. Then, (5.59) becomes

$$\left(\mathcal{D}_t + \frac{V\mu}{\lambda+\mu}\mathcal{D}_x\right) A + \frac{\mu}{\lambda+\mu}\left(\gamma A^2 - \delta C\right) = 0. \quad (5.60)$$

We also know that C satisfies (5.54), which we can write as

$$(\mathcal{D}_t + V\mathcal{D}_x)\, C - \gamma A^2 + \delta C = 0, \tag{5.61}$$

where we have assumed that $\theta = \frac{1}{2}$ and so $\mathcal{M}_t = 1$. These are the key equations as mentioned in the Introduction to this Chapter. Equation (5.61) shows us that for a fixed x, C will move faster than A with speed V. This drives A along with the boundary data for A. Equation (5.60) shows us that generally A travels at speed $\frac{V\mu}{\lambda+\mu}$ with slow damping by the C term.

However, as we have observed in the previous Section this is not always the case. The left two sets of plots in Figure 5-2 show results for fixed $K = 5$ but γ and δ are different sizes. The right plots show a similar situation, but with $K = 1/5$. When $K = 5$ all the concentrations move faster as γ and δ increase. When $K = 1/5$, on the other hand, a and b always move at the reduced speed and c slows down as γ and μ increase. However, c now has two peaks when these parameters are small, one of which is moving at speed V and the other one at a much slower speed. These features will be explained using the modified equation analysis.

Lighthill & Whitham (1955) considered a model to describe flood movement in long rivers which consisted of a pair of coupled first order nonlinear hyperbolic equations. They derived a second order equation in only one variable and linearised to perform analysis on the model. We follow a similar procedure for (5.60) and (5.61) to obtain a second order equation for A only and then linearise the A^2 term. Suppose we separately apply $\mathcal{D}_t$ and $\mathcal{D}_x$ to (5.60). Then

$$\mathcal{D}_t^2 A + V'\mathcal{D}_x\mathcal{D}_t A + \tfrac{\mu}{\lambda+\mu}\left[\gamma\mathcal{D}_t(A^2) - \delta\mathcal{D}_t C\right] = 0 \tag{5.62}$$

$$\mathcal{D}_x D_t A + V'\mathcal{D}_x^2 A + \tfrac{\mu}{\lambda+\mu}\left[\gamma\mathcal{D}_x(A^2) - \delta\mathcal{D}_x C\right] = 0. \tag{5.63}$$

Also, multiplying (5.61) by $\frac{\mu}{\lambda+\mu}$ and adding the result to (5.60) gives

$$\mathcal{D}_t A + V'\mathcal{D}_x A + \frac{\mu}{\lambda+\mu}\left(\mathcal{D}_t C + V\mathcal{D}_x C\right) = 0. \tag{5.64}$$

Then, multiplying (5.63) by V and adding the result to (5.62) leads to

$$\mathcal{D}_t^2 A + (V'+V)\mathcal{D}_x\mathcal{D}_t A + VV'\mathcal{D}_x^2 A + \frac{\gamma\mu}{\lambda+\mu}\left[\mathcal{D}_t(A^2) + V\mathcal{D}_x(A^2)\right] - \frac{\delta\mu}{\lambda+\mu}\left(\mathcal{D}_t C + V\mathcal{D}_x C\right) = 0. \tag{5.65}$$

Finally, the last term in (5.65) can be replaced using (5.64) and so we obtain a second order equation entirely in terms of A

$$\mathcal{D}_t^2 A + (V'+V)\mathcal{D}_x\mathcal{D}_t A + VV'\mathcal{D}_x^2 A + \mathcal{D}_t\left[\delta A + \frac{\gamma\mu}{\lambda+\mu}A^2\right] + V'\mathcal{D}_x\left[\delta A + \gamma A^2\right] = 0. \tag{5.66}$$

This is now linearised by setting $A = A_0 + \bar{A}$ and retaining only the first-order terms in $\bar{A}$. Then

$$\mathcal{D}_t^2\bar{A} + (V' + V)\mathcal{D}_x\mathcal{D}_t\bar{A} + VV'\mathcal{D}_x^2\bar{A} + \mathcal{D}_t\left[\delta\bar{A} + \frac{2\gamma\mu A_0}{\lambda+\mu}\bar{A}\right] + V'\mathcal{D}_x[\delta\bar{A} + 2\gamma A_0\bar{A}] = 0. \tag{5.67}$$

Equation (5.67) leads to a quadratic equation for the operator $\mathcal{D}_t$, i.e.

$$\mathcal{D}_t^2 + [\omega + \xi\mathcal{D}_x]\mathcal{D}_t + \eta\mathcal{D}_x^2 + \nu\mathcal{D}_x = 0, \tag{5.68}$$

where

$$\omega := \delta + \frac{2\gamma\mu A_0}{\lambda+\mu}, \quad \xi := \frac{V(\lambda+2\mu)}{\lambda+\mu}, \quad \eta := \frac{V^2\mu}{\lambda+\mu}, \quad \nu := \frac{V\mu}{\lambda+\mu}[\delta + 2\gamma A_0]. \tag{5.69}$$

On solving (5.68) we obtain two roots; these can be expanded to give expansions in increasing powers of $\mathcal{D}_x$ (exactly as for the Linear Model in Chapter 3). We wish to explain the phenomenon of varying speeds observed in the figures in Section 5.3.1 and so only consider up to $\mathcal{D}_x$ terms (since we are not interested in any diffusive effects). So, the positive and negative roots are

$$\mathcal{D}_t = -\frac{\nu}{\omega}\mathcal{D}_x + \ldots, \qquad \mathcal{D}_t = -\omega - \left(\xi - \frac{\nu}{\omega}\right)\mathcal{D}_x + \ldots, \tag{5.70}$$

respectively. When ω is large the second root quickly tends to zero and will not be observed. $\bar{A}$ can be applied to these equations for the differential operators in (5.70). Hence, truncating the expansions after the first order terms gives

$$\bar{A}_t + \frac{\nu}{\omega}\bar{A}_x = 0, \tag{5.71}$$

for the positive root, and

$$\bar{A}_t + \left(\xi - \frac{\nu}{\omega}\right)\bar{A}_x = -\omega\bar{A}, \tag{5.72}$$

for the negative root, where

$$\frac{\nu}{\omega} = \frac{V\mu(1+2KA_0)}{\lambda+\mu+2K\mu A_0}, \qquad \xi - \frac{\nu}{\omega} = V\left(\frac{\lambda+\mu+2K\mu\left(\frac{\mu}{\lambda+\mu}\right)A_0}{\lambda+\mu+2K\mu A_0}\right), \tag{5.73}$$

and ω is given in (5.69).

In Section 5.2.1 we showed that the Equilibrium model, in terms of a, can be written as $a_t + F(a)a_x = 0$, where $F(a)$ $(= f'(a))$ comes directly from (5.39). Comparing this with (5.71) we see that the two equations are practically identical (except that $F(a)$ has now been linearised). So, when ω is large (which corresponds to large δ and μ), the

	ν/ω	ω	$\xi - \nu/\omega$
$\gamma = 5,\ \delta = 1,\ A_0 = 0.1$	0.1818	1.10	1.089
$\gamma = 50,\ \delta = 10,\ A_0 = 0.2$	0.250	12.0	1.176
$\gamma = 1,\ \delta = 5,\ A_0 = 0.1$	0.1036	5.020	1.0036
$\gamma = 10,\ \delta = 50,\ A_0 = 0.12$	0.1043	50.24	0.9957

Table 5.1: *A Table showing the values of the coefficients from (5.71) and (5.72) using the parameters from Figure 5-2 ($\lambda = 90$ and $\mu = 10$ in all cases).*

negative root will decay to zero very quickly and then (5.71) will accurately describe the Flushing-through Model. However, when δ is not large the second root will have an impact and this explains why we observe an extra peak in Figure 5-2. Also, this analysis shows that the Equilibrium model can only be accurate when δ and μ are large, confirming our findings in Section 5.3.3.

Table 5.1 shows the values of the coefficients in the modified equation expansions using the parameter values from Figure 5-2. We have taken A_0 to be the maximum value of the numerical solution A at the fixed distance $x = 1$. In the top left plots (which correspond to the first row of values) ω is small and so the negative root will be very significant. This explains why the pulses for a and b are very wide and are not moving at either of the speeds ν/ω or $\xi - \nu/\omega$. The bottom left plots (corresponding to the second row of values) show the three pulses moving at the same speed ν/ω (which is not the reduced speed) with similar shape. This is because ω is very large and so (5.72) is negligible.

In both sets in the right plots in Figure 5-2 (corresponding to the bottom two rows of the table) the pulses a and b move with speed ν/ω. The parameter ω is fairly large and so the negative root does not have much affect on a and b. However, the pulse for c has an extra spike which is directly caused by this root: it moves at roughly the speed $\xi - \nu/\omega$ as shown in the third row. This disappears in the bottom right set of plots because ω is very large.

In conclusion, the modified equation analysis has allowed us to explain some features of the model which were not easy to identify from examining the original formulation. The analysis is valid for all sizes of the parameters, although we have restricted λ and μ large to predict features of the Flushing-through Model; especially the extra peak for c and the fact that a and c can move at different speeds (which cannot be explained using the Improved-equilibrium model as this results in a single convection-diffusion equation).

Chapter 6

Hyperbolic conservation laws with source terms in 2D

6.1 Introduction

In previous Chapters we have studied only 1D model problems describing the transport of chemical pollutants in groundwater flow. However, it is more realistic to extend these models to 2D; we will focus attention on situations where some chemical pollutants are inserted into the water at a certain height and spread both vertically and horizontally through the water in time. A specific example is given in (Walter et al. 1994*a*) and (Walter, Frind, Blowes, Ptacek & Molson 1994*b*) where the authors consider the mobility of potentially toxic dissolved metals discharged from a mine tailings source into an aquifer (the 2D flow field is shown in Figure 6-1). We will simplify this situation, to make it more comparable with reactive transport problems, by supposing the tailings are a pulse of chemical pollutants which enter the groundwater at some height Z above the ground. However, we will use the flux conditions on the boundary as specified in (Walter et al. 1994*b*, page 3151) to give realistic velocity profiles.

In this Chapter the weighted box-trap scheme is applied to these types of problems. We begin by describing the box scheme in 2D for a hyperbolic first order system. In Section 6.2 some simple model problems in 2D are considered where the exact solution can be found using the method of characteristics. Unlike the 1D problems studied in Chapter 3, the oscillations cannot be eliminated by choosing $\theta > \frac{1}{2}$. This is due to the fluxes being non-constant and so, in Section 6.3, we return to considering hyperbolic equations in 1D, but with variable speeds, to gain insight into this more realistic case. A modified equation analysis is used to predict optimal choices of the mesh, and reduce potential oscillations. This involves using a variable spatial step length which eliminates the dispersion term in the modified equation expansion. The modified equation analysis has proved throughout this Thesis to be a very valuable tool in describing both

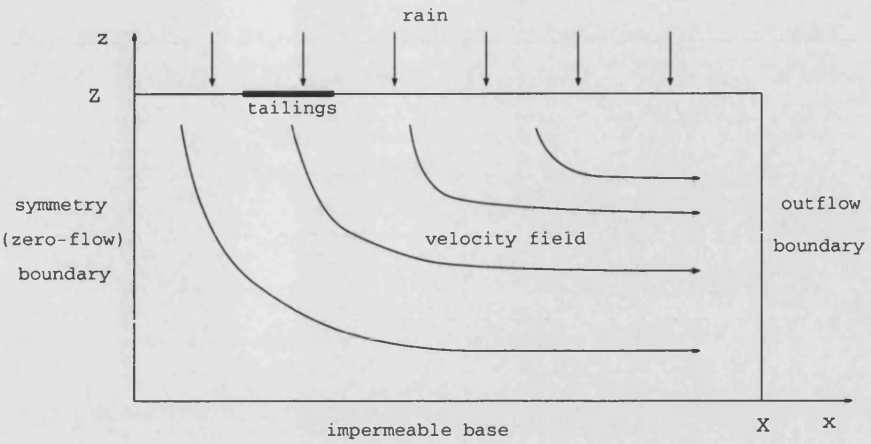

Figure 6-1: *2D cross-sectional plane with the velocity field and boundary conditions.*

numerical and analytical features of reactive transport models.

We return to the 2D problem in Section 6.4 where the ideas from the analysis of 1D problems are applied. This situation is more complicated because the fluxes are functions of two variables. However, we are able to show that a variable mesh works well in many cases, although a constant mesh is sometimes sufficient, when the fluxes do not change much.

6.1.1 The weighted box scheme in 2D

Consider the general problem as defined in the Section 1.3.4 of the Introduction, i.e.

$$(\mathbf{u_1})_t + \nabla \cdot (\nabla\phi \mathbf{u_1}) = \mathbf{S_1}(\mathbf{u_1}, \mathbf{u_2}) \tag{6.1}$$

$$(\mathbf{u_2})_t = \mathbf{S_2}(\mathbf{u_1}, \mathbf{u_2}), \tag{6.2}$$

where $\mathbf{u_1} = \mathbf{u_1}(x, z, t)$, $\mathbf{u_2} = \mathbf{u_2}(x, z, t)$ and we consider the region $\{(x, z) \mid 0 \leq x \leq X,\ 0 \leq z \leq Z\}$ with appropriate conditions on the boundaries. Following the work in the previous Chapters we wish to apply the weighted box scheme to (6.1) and the trapezoidal scheme to (6.2). The latter is a straightforward extension of the 1D case and so we concentrate on (6.1) and write as

$$\mathbf{u}_t + \nabla \cdot (\nabla\phi \mathbf{u}) = \mathbf{S}(\mathbf{u}). \tag{6.3}$$

In practical applications the flow will often be incompressible and so

$$\nabla^2 \phi(x, z) = 0, \tag{6.4}$$

which means that

$$\nabla \cdot (\nabla \phi \mathbf{u}) = (\nabla^2 \phi) \cdot \mathbf{u} + (\nabla \phi) \cdot (\nabla \mathbf{u}) = (\nabla \phi) \cdot (\nabla \mathbf{u}).$$

Hence we could replace (6.3) (which is in divergence form) by the equivalent form

$$\mathbf{u_t} + (\nabla \phi) \cdot (\nabla \mathbf{u}) = \mathbf{S}(\mathbf{u}). \tag{6.5}$$

However, the box scheme discretisation of the spatial derivatives in (6.3) and (6.5) will take different forms: since $\nabla \phi$ appears under the divergence sign in the former this would be included in the approximation of this derivative. In the latter, $\nabla \phi$ would be approximated independently of $\nabla \cdot \mathbf{u}$. We have performed numerical simulations on a simple example to compare both formulations. Consider

$$u_t + (1+x)u_x - (1+z)u_z = 0 \quad \text{and} \quad u_t + \big((1+x)u\big)_x - \big((1+z)u\big)_z = 0. \tag{6.6}$$

Although not reproduced here, the weighted box scheme applied to the two formulations in (6.6) produced very similar results for a step function boundary condition. On the basis of this experiment, we will consider the formulation given in (6.5) as the discretised equations are simpler to describe. In Section 6.2.2 we will explicitly show how poor the weighted box scheme behaves for the first example in (6.6). Note that in 1D the equation (6.4) is equivalent to requiring $\phi''(x) = 0$ and so the flux must be constant. This explains why we have only considered constant speeds in our simple 1D models. However, we will consider 1D problems with non-constant speed in Section 6.3 to understand how the box scheme copes with this phenomenon.

The discretisation of (6.5) using the weighted box scheme is now described. For convenience we set $\phi_x = a(x,z)$ and $\phi_z = b(x,z)$ and so (6.5) can be written as

$$\mathbf{u}_t + a(x,z)\mathbf{u}_x + b(x,z)\mathbf{u}_z = \mathbf{S}(\mathbf{u}), \tag{6.7}$$

In terms of difference operators the basic box scheme applied to (6.7) is given by

$$\mu_x\mu_z\delta_t \mathbf{U}^{n+\frac{1}{2}}_{i+\frac{1}{2},j+\frac{1}{2}} + \nu_x \bar{A}\mu_z\mu_t\delta_x \mathbf{U}^{n+\frac{1}{2}}_{i+\frac{1}{2},j+\frac{1}{2}} + \nu_z \bar{B}\mu_x\mu_t\delta_z \mathbf{U}^{n+\frac{1}{2}}_{i+\frac{1}{2},j+\frac{1}{2}} = \Delta t \mu_t\mu_x\mu_z \mathbf{S}^{n+\frac{1}{2}}_{i+\frac{1}{2},j+\frac{1}{2}}, \tag{6.8}$$

where $\nu_x := \Delta t/\Delta x$, $\nu_z := \Delta t/\Delta z$ are the mesh ratios and $\bar{A}$ and $\bar{B}$ are taken to be the cell averages

$$\bar{A} \equiv \mu_x\mu_z A_{i+\frac{1}{2},j+\frac{1}{2}} = \tfrac{1}{4}\left(A_{i+1,j+1} + A_{i,j+1} + A_{i+1,j} + A_{i,j}\right) \tag{6.9}$$

$$\bar{B} \equiv \mu_x\mu_z B_{i+\frac{1}{2},j+\frac{1}{2}} = \tfrac{1}{4}\left(B_{i+1,j+1} + B_{i,j+1} + B_{i+1,j} + B_{i,j}\right). \tag{6.10}$$

Also set

$$P := \nu_x A_{i+\frac{1}{2},j+\frac{1}{2}}, \qquad Q := \nu_z B_{i+\frac{1}{2},j+\frac{1}{2}}, \tag{6.11}$$

then (6.8) becomes

$$\begin{aligned}&\mu_x\mu_z \mathbf{U}^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}} + \tfrac{1}{2}P\mu_z\delta_x \mathbf{U}^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}} + \tfrac{1}{2}Q\mu_x\delta_z \mathbf{U}^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}} - \tfrac{1}{2}\triangle t\mu_x\mu_z \mathbf{S}^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}}\\ &= \mu_x\mu_z \mathbf{U}^{n}_{i+\frac{1}{2},j+\frac{1}{2}} - \tfrac{1}{2}P\mu_z\delta_x \mathbf{U}^{n}_{i+\frac{1}{2},j+\frac{1}{2}} - \tfrac{1}{2}Q\mu_x\delta_z \mathbf{U}^{n}_{i+\frac{1}{2},j+\frac{1}{2}} + \tfrac{1}{2}\triangle t\mu_x\mu_z \mathbf{S}^{n}_{i+\frac{1}{2},j+\frac{1}{2}}.\end{aligned} \tag{6.12}$$

Now

$$\begin{aligned}\mu_x\mu_z \mathbf{U}^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}} &= \tfrac{1}{4}\left(\mathbf{U}^{n+1}_{i+1,j+1} + \mathbf{U}^{n+1}_{i,j+1} + \mathbf{U}^{n+1}_{i+1,j} + \mathbf{U}^{n+1}_{i,j}\right)\\ \mu_z\delta_x \mathbf{U}^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}} &= \tfrac{1}{2}\left(\mathbf{U}^{n+1}_{i+1,j+1} + \mathbf{U}^{n+1}_{i+1,j}\right) - \tfrac{1}{2}\left(\mathbf{U}^{n+1}_{i,j+1} + \mathbf{U}^{n+1}_{i,j}\right)\\ \mu_x\delta_z \mathbf{U}^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}} &= \tfrac{1}{2}\left(\mathbf{U}^{n+1}_{i+1,j+1} + \mathbf{U}^{n+1}_{i,j+1}\right) - \tfrac{1}{2}\left(\mathbf{U}^{n+1}_{i+1,j} + \mathbf{U}^{n+1}_{i,j}\right),\end{aligned}$$

and so (6.12) can be written as

$$\begin{aligned}&\left[(I+P+Q)\mathbf{U}^{n+1} - \tfrac{1}{2}\triangle t\mathbf{S}^{n+1}\right]_{i+1,j+1} + \left[(I-P+Q)\mathbf{U}^{n+1} - \tfrac{1}{2}\triangle t\mathbf{S}^{n+1}\right]_{i,j+1}\\ &+ \left[(I+P-Q)\mathbf{U}^{n+1} - \tfrac{1}{2}\triangle t\mathbf{S}^{n+1}\right]_{i+1,j} + \left[(I-P-Q)\mathbf{U}^{n+1} - \tfrac{1}{2}\triangle t\mathbf{S}^{n+1}\right]_{i,j} = 4\mathbf{R}^{\mathbf{n}}.\end{aligned} \tag{6.13}$$

where $\mathbf{R}^{\mathbf{n}}$ consists of known values at level n and is given by

$$\begin{aligned}\mathbf{R}^{\mathbf{n}} = &\tfrac{1}{4}\left[(I-P-Q)\mathbf{U}^{n} + \tfrac{1}{2}\triangle t\mathbf{S}^{n}\right]_{i+1,j+1} + \tfrac{1}{4}\left[(I+P-Q)\mathbf{U}^{n} + \tfrac{1}{2}\triangle t\mathbf{S}^{n}\right]_{i,j+1}\\ &+\tfrac{1}{4}\left[(I-P+Q)\mathbf{U}^{n} + \tfrac{1}{2}\triangle t\mathbf{S}^{n}\right]_{i+1,j} + \tfrac{1}{4}\left[(I+P+Q)\mathbf{U}^{n} + \tfrac{1}{2}\triangle t\mathbf{S}^{n}\right]_{i,j}.\end{aligned}$$

Consider the left set of boxes in Figure 6-2. We have four different cases depending on whether the velocity fields are positive or negative (we assume that both are non-zero).

- *The pp-case.* If $A > 0$ and $B > 0$ then the unknown value is $\mathbf{U}^{n+1}_{i+1,j+1}$. Hence $P > 0$ and $Q > 0$ and so $P = |P|$ and $Q = |Q|$.
- *The pm-case.* If $A > 0$ and $B < 0$ then the unknown value is $\mathbf{U}^{n+1}_{i+1,j}$. Hence $P > 0$ and $Q < 0$ and so $P = |P|$ and $Q = -|Q|$.
- *The mp-case.* If $A < 0$ and $B > 0$ then the unknown value is $\mathbf{U}^{n+1}_{i,j+1}$. Hence $P < 0$ and $Q > 0$ and so $P = -|P|$ and $Q = |Q|$.
- *The mm-case.* If $A < 0$ and $B < 0$ then the unknown value is $\mathbf{U}^{n+1}_{i,j}$. Hence $P < 0$ and $Q < 0$ and so $P = -|P|$ and $Q = -|Q|$.

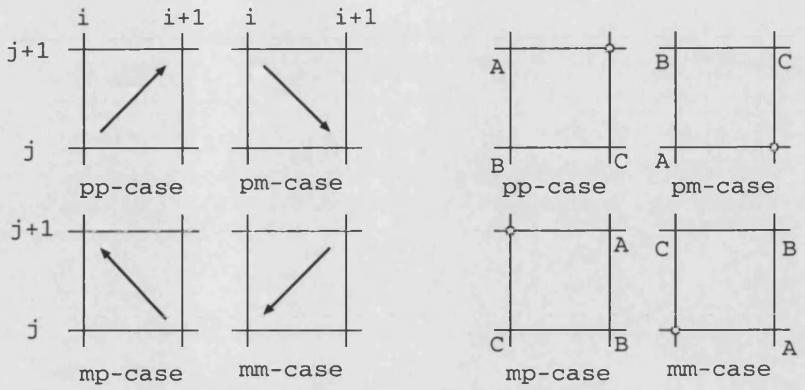

Figure 6-2: *The left set of boxes shows the four different cases for the velocity fields and the right set of boxes shows the direction to solve for the four different cases.*

Suppose that, in labelling the points A, B and C, from the unknown point we always go to the node at the same j^{th} level and then continue round to the other points (see the right set of boxes in Figure 6-2 where the unknown value is denoted by an open circle). This means in general we can write

$$\begin{aligned}\left[(I+|P|+|Q|)\mathbf{U}^{n+1}-\tfrac{1}{2}\triangle t\mathbf{S}^{n+1}\right]_{\text{unknown}} &= -\left[(I-|P|+|Q|)\mathbf{U}^{n+1}-\tfrac{1}{2}\triangle t\mathbf{S}^{n+1}\right]_{A}\\ -\left[(I-|P|-|Q|)\mathbf{U}^{n+1}-\tfrac{1}{2}\triangle t\mathbf{S}^{n+1}\right]_{B} &- \left[(I+|P|-|Q|)\mathbf{U}^{n+1}-\tfrac{1}{2}\triangle t\mathbf{S}^{n+1}\right]_{C}+4\mathbf{R}^{\mathbf{n}}.\end{aligned} \tag{6.14}$$

In Chapter 3 the weighted-box scheme was applied to linear problems in 1D; for systems in 2D, such as (6.7), this corresponds to replacing the difference operator μ_t on the left hand side of (6.8) with θ_t (defined in (3.62)). Then (6.13) is

$$\begin{aligned}&\left[(I+2\theta(P+Q))\mathbf{U}^{n+1}-\tfrac{1}{2}\triangle t\mathbf{S}^{n+1}\right]_{i+1,j+1}+\left[(I-2\theta(P-Q))\mathbf{U}^{n+1}-\tfrac{1}{2}\triangle t\mathbf{S}^{n+1}\right]_{i,j+1}\\ &+\left[(I+2\theta(P-Q))\mathbf{U}^{n+1}-\tfrac{1}{2}\triangle t\mathbf{S}^{n+1}\right]_{i+1,j}+\left[(I-2\theta(P+Q))\mathbf{U}^{n+1}-\tfrac{1}{2}\triangle t\mathbf{S}^{n+1}\right]_{i,j}\\ &=\left[(I-2(1-\theta)(P+Q))\mathbf{U}^{n}+\tfrac{1}{2}\triangle t\mathbf{S}^{n}\right]_{i+1,j+1}+\left[(I+2(1-\theta)(P-Q))\mathbf{U}^{n}+\tfrac{1}{2}\triangle t\mathbf{S}^{n}\right]_{i,j+1}\\ &+\left[(I-2(1-\theta)(P-Q))\mathbf{U}^{n}+\tfrac{1}{2}\triangle t\mathbf{S}^{n}\right]_{i+1,j}+\left[(I+2(1-\theta)(P+Q))\mathbf{U}^{n}+\tfrac{1}{2}\triangle t\mathbf{S}^{n}\right]_{i,j}.\end{aligned} \tag{6.15}$$

In the mine tailings problem, data is prescribed on the boundaries $x = 0$ and $z = Z$ for all time. So, using the terminology above, we are in the *pm-case* because the velocity field travels from the top left to bottom right (see Figure 6-1).

6.2 Simple problems in 2D

6.2.1 A conservation law with constant velocity flux

Suppose we have a simple conservation law with constant fluxes in 2D, say

$$u_t + u_x - u_z = 0, \tag{6.16}$$

where $0 \leq t \leq T$, $0 \leq x \leq X$ and $0 \leq z \leq Z$. We are again in the *pm-case* and so need to specify boundary data on the $x = 0$ and $z = Z$ boundaries so we can solve to find the unknown value $U_{i+1,j}^{n+1}$ for $i = 0, \ldots, I-1$ and $j = J-1, \ldots, 0$ at each level n (for $n = 0, \ldots, N-1$). In the notation from the previous Section P and Q are simply

$$P = \nu_x, \qquad Q = -\nu_z,$$

and the source term $\mathbf{S}$ is zero. We now follow (McOwen 1995, pages 15-18) where the method of characteristics is described for scalar equations with more than two variables. If we specify $u(x, Z, t) = f(x, t)$ then the solution is

$$u(x, z, t) = f(x - Z + z, t - Z + z). \tag{6.17}$$

Suppose

$$u(0, z, t) = 0, \qquad 0 \leq z \leq Z, \quad 0 \leq t \leq T, \tag{6.18}$$

$$u(x, Z, t) = \begin{cases} g(x)e^{-10(t-T_s)^2}, & \text{if } 0 \leq t \leq T_s \\ g(x), & \text{otherwise,} \end{cases} \tag{6.19}$$

for $0 \leq x \leq X$, where $g(x)$ is a short injection of a chemical pulse. The second condition says there is a slow input of chemicals up until some time $t = T_s$. From this time onwards the condition on the $z = Z$ boundary is simply the pulse itself. We always assume $T > T_s$ because, if $T_s = 0$, there is an immediate switch on which can cause extra oscillations. The condition (6.19) defines $f(x, t)$ and so (6.17) becomes

$$u(x, z, t) = \begin{cases} g(x - Z + z)e^{-10(t-Z+z-T_s)^2}, & \text{if } 0 \leq t - Z + z \leq T_s \\ g(x - Z + z), & \text{otherwise,} \end{cases} \tag{6.20}$$

for $0 < x \leq X$ and $0 \leq z \leq Z$ where (6.18) holds at $x = 0$. Consider a pulse

$$g(x) = \begin{cases} 0, & x \leq \tau \\ \sin^2\left(\frac{\alpha\pi}{2}(x - \tau)\right), & \tau < x < \frac{1}{\alpha} + \tau \\ 1, & \frac{1}{\alpha} + \tau \leq x \leq \beta - \frac{1}{\alpha} + \tau \\ \sin^2\left(\frac{\alpha\pi}{2}(\beta + \tau - x)\right), & \beta - \frac{1}{\alpha} + \tau \leq x \leq \beta + \tau \\ 0, & x > \beta + \tau, \end{cases} \tag{6.21}$$

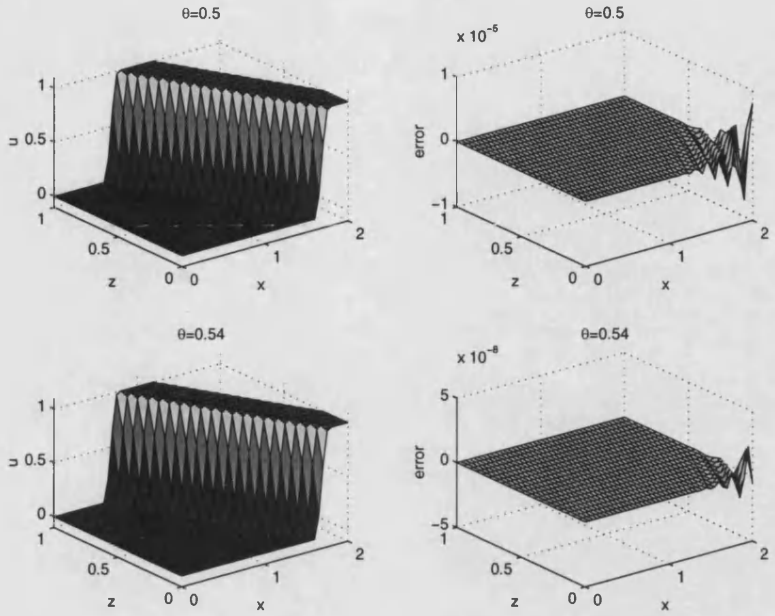

Figure 6-3: *The left plots show the weighted box scheme applied to (6.16) with* $\theta = 0.5$ *and* $\theta = 0.54$ *respectively. The right plots show the error. These are plotted at time* $t = 3$ *with* $\alpha = 6$, $\tau = 0.6$, $\beta = 0.8$, $T_s = 1$, $\triangle t = 0.04$ *and* $\nu_x = \nu_z = 0.8$.

which is smoother than a square pulse. We call this the *sin-squared pulse.* Hence for small α the pulse is smooth and, as α increases, this becomes more like a square pulse. In Figure 6-3 the weighted box scheme is applied to (6.16) (with $\theta = 0.5$ in the top plots and $\theta = 0.54$ in the bottom plots). There are some very small oscillations; this is not clear from the left plots but can be seen by examining the errors on the right: the top right is $O(10^{-5})$ and the bottom right is $O(10^{-8})$. They are reduced by using the weighted box scheme with $\theta = 0.5 + \triangle t = 0.54$. However, we are interested in problems with variable flux functions and so we now study a more interesting example.

6.2.2 The poor performance of the weighted box scheme

Suppose we have the following linearised conservation law:

$$u_t + (1 + x)u_x - (1 + z)u_z = 0, \tag{6.22}$$

for $0 \leq t \leq T$, $0 \leq x \leq X$ and $0 \leq z \leq Z$. In this case

$$P = \tfrac{1}{2}(2 + x_i + x_{i+1})\nu_x, \quad Q = -\tfrac{1}{2}(2 + z_j + z_{j+1})\nu_z,$$

where $x_i = i\triangle x$ and $z_j = j\triangle z$ for $i = 0, \ldots, I$ and $j = 0, \ldots, J$ respectively. Following the procedure of Section 6.2.1 we can solve (6.22) with $u(x, Z, t) = f(x, t)$ using the

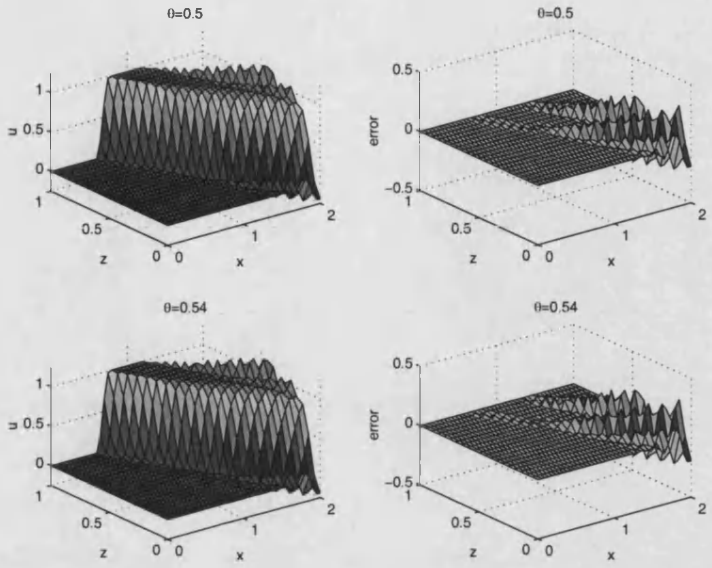

Figure 6-4: *The left plots show the weighted box scheme applied to (6.22) with* $\theta = 0.5$ *and* $\theta = 0.54$ *respectively. The right plots show the error. These are plotted at time* $t = 3$ *with* $\alpha = 6$, $\tau = 0.6$, $\beta = 0.8$, $T_s = 1$, $\triangle t = 0.04$ *and* $\nu_x = \nu_z = 0.8$.

method of characteristics. Then the solution is found to be (assuming (6.18) and (6.19) are specified on the boundaries)

$$u(x,z,t) = \begin{cases} g\left(\frac{(1+x)(1+z)}{1+Z}\right) e^{-10\left[t+\ln\left(\frac{1+z}{1+Z}\right)-T_s\right]^2} & \text{if } 0 \le t + \ln\left(\frac{1+z}{1+Z}\right) \le T_s \\ g\left(\frac{(1+x)(1+z)}{1+Z}\right) & \text{otherwise,} \end{cases} \tag{6.23}$$

for $0 < x \le X$ and $0 \le z \le Z$ with (6.18) holding at $x = 0$. In Figure 6-4 we show the weighted box scheme applied to the same problem. The oscillations here are very severe and are not reduced by taking $\theta > \frac{1}{2}$. Indeed, on comparing the errors in the right plots, it seems that the weighted box scheme is no better than the box scheme.

This Figure illustrates that using the weighted box scheme is not sufficient to eliminate the oscillations when the fluxes are not constant. We need to find another way to reduce the observed oscillations in these more complicated problems.

6.3 Hyperbolic equations in 1D with variable speed

In an attempt to understand and consequently control the oscillations observed when applying the box scheme to problems with a nonlinear flux function, first consider the modified equation analysis of the following 1D equation:

$$u_t + a(x)u_x = -\lambda u, \tag{6.24}$$

where $\lambda > 0$ is assumed constant. The weighted box scheme applied to (6.24) is

$$\mu_x \delta_t U_{i+1/2}^{n+1/2} + \nu \bar{A} \theta_t \delta_x U_{i+1/2}^{n+1/2} = -\lambda' \mu_x \mu_t U_{i+1/2}^{n+1/2}, \tag{6.25}$$

where, since a is a function of x, it has been approximated by $\bar{A} := \mu_x a_{i+1/2}$. We can expand these finite difference operators (as defined in Section 6.4.1 of Chapter 3) to give the modified equation expansion of (6.25). Now

$$\mu_x a(x) = \left(1 + \tfrac{1}{8}\triangle x^2 \partial_x^2 + \ldots\right) a(x) = a(x) + \tfrac{1}{8}\triangle x^2 a''(x) + \ldots.$$

Following the procedure of Chapter 3, the discretised equations in (6.25) can be expanded and an expression found for U_t in terms of x derivatives. This is similar to our discussions in Chapter 3 except that the speed now depends on x. We know that there are two key terms in the modified equation expansion. Firstly, the dispersion term which relates the cell size to the velocity; if this is set to zero then the oscillations can be controlled. Secondly, the diffusion term which, if set to zero, will minimise the numerical diffusion in the scheme. In 1D both terms can be set to zero in the weighted box scheme by choice of θ and $\triangle x$. When $a(x)$ is variable the situation is the same but we will need to vary $\triangle x$ in order for the dispersion term to be set to zero.

After some manipulation the modified equation expansion of (6.25) is found to be

$$\begin{aligned} U_t \;=\; & -\lambda\left[1 + \tfrac{1}{12}\triangle t^2 \lambda^2\right] U - a(x)\left[1 + \tfrac{1}{4}\lambda^2 \triangle t^2 - \lambda\left(\theta - \tfrac{1}{2}\right)\triangle t\right] U_x \\ & + \triangle t a(x) a'(x)\left\{\left(\theta - \tfrac{1}{2}\right) - \lambda\left(\theta - \tfrac{1}{2}\right)^2 \triangle t - \tfrac{1}{4}\lambda \triangle t\right\} U_x \\ & - \tfrac{1}{12}\triangle t^2 a(x)\left[a'(x)^2 + a(x)a''(x)\right]\left[1 + 12\left(\theta - \tfrac{1}{2}\right)^2\right] U_x \\ & + \tfrac{1}{4}a'(x)\left\{\triangle x^2 - \triangle t^2 a(x)^2 - 12\left(\theta - \tfrac{1}{2}\right)^2 \triangle t^2 a(x)^2\right\} U_{xx} \\ & + \triangle t a(x)^2 \left\{\left(\theta - \tfrac{1}{2}\right) - \lambda\left(\theta - \tfrac{1}{2}\right)^2 \triangle t - \tfrac{1}{4}\lambda \triangle t\right\} U_{xx} \\ & + \tfrac{1}{12}a(x)\left\{\triangle x^2 - \triangle t^2 a(x)^2 - 12\left(\theta - \tfrac{1}{2}\right)^2 \triangle t^2 a(x)^2\right\} U_{xxx} + \ldots. \end{aligned} \tag{6.26}$$

Figure 6-5 shows plots of the numerical solution (dashed line) compared with the exact solution (solid line, again easily found using the method of characteristics) with $a(x) = 1 + x$, $\lambda = 1$ and $\theta = \frac{1}{2}$ for various initial conditions. As expected, the oscillations disappear when the pulse becomes more smooth. However, our aim is to find a numerical scheme which is robust and accurate for all types of initial data and so we want to reduce the oscillations for the top two cases in Figure 6-5. Previously, this has been successfully achieved using the weighted box scheme. Figure 6-6 shows the results with $\theta = 0.5 + \triangle t = 0.52$ for the same problem. Whilst reducing the oscillations slightly, the weighted box scheme introduces a significantly large amount of damping. This is much more noticeable for non-constant a and so weighting the box scheme does

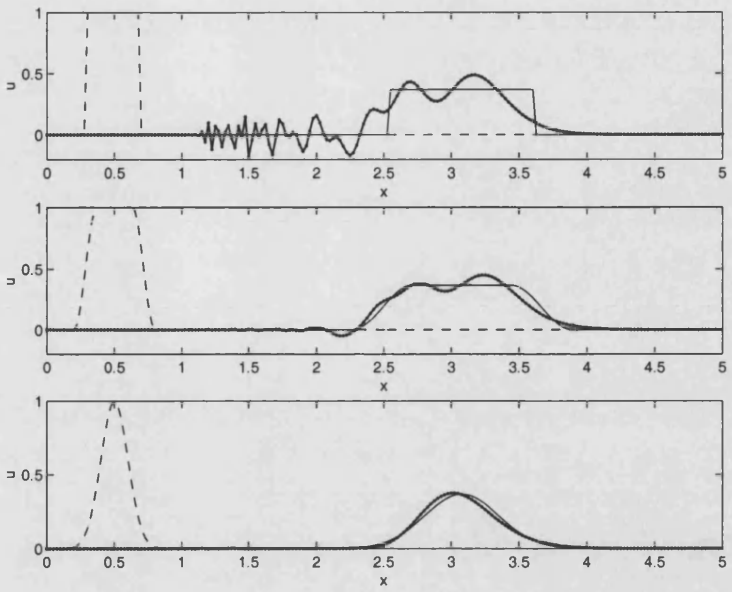

Figure 6-5: *Plots of the box scheme (with $\theta = \frac{1}{2}$) applied to (6.24) with $a(x) = 1 + x$ and $\lambda = 1$ at time $t = 1$ for a square pulse (top plot), a sin-squared pulse (middle plot) and a Gaussian pulse (bottom plot). The dashed line denotes the initial condition, the dots joined by an unbroken line denotes the numerical solution and the thin unbroken line denotes the exact solution, with $\triangle t = 0.02$ and $\triangle x = 0.025$.*

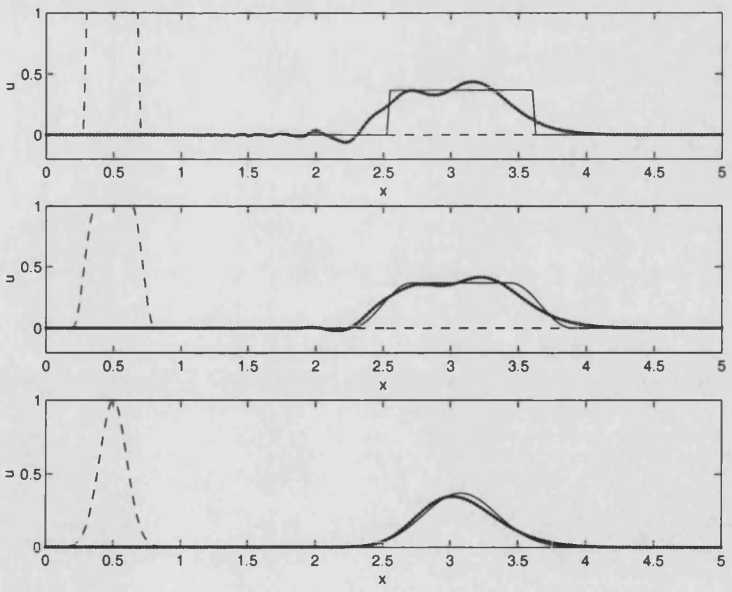

Figure 6-6: *Plots of the weighted box scheme ($\theta = 0.52$) applied to (6.24) with $a(x) = 1 + x$ and $\lambda = 1$ at time $t = 1$ for a square pulse (top plot), a sin-squared pulse (middle plot) and a Gaussian pulse (bottom plot). The dashed line denotes the initial condition, the dots joined by an unbroken line denotes the numerical solution and the thin unbroken line denotes the exact solution, with $\triangle t = 0.02$ and $\triangle x = 0.025$.*

not seem to be sufficient here.

As discussed above we can set the dispersion term to zero in (6.26). Since $a(x)$ is not constant this means we choose the spatial step length to depend on x, i.e.

$$\Delta x = \Delta t a(x)\sqrt{1 + 12\left(\theta - \tfrac{1}{2}\right)^2}. \tag{6.27}$$

It makes sense to fix Δt and vary Δx since a is a function of x only and we have non-zero data on the x boundary. Also, by observing the U_{xx} terms in (6.26) we see that the first is eliminated with the above choice of Δx; the second can also be set to zero if θ is chosen to satisfy

$$\left(\theta - \tfrac{1}{2}\right) - \lambda\left(\theta - \tfrac{1}{2}\right)^2 \Delta t - \tfrac{1}{4}\lambda\Delta t = 0, \tag{6.28}$$

Note that if $\lambda = 0$ then (6.28) simply becomes $\theta = \frac{1}{2}$. If $\lambda\Delta t > 1$ then the discriminant in (6.28) is negative and so the U_{xx} term cannot be eliminated with a positive θ. We would then have to choose θ to minimise this term. If $\lambda\Delta t \leq 1$ then (6.28), which is a quadratic equation for $\theta - \frac{1}{2}$, can be solved and we take the negative square root since $\left(\theta - \frac{1}{2}\right)_+ \to \infty$ as $\Delta t \to 0$ whereas $\left(\theta - \frac{1}{2}\right)_- \to 0$ as $\Delta t \to 0$. Since $\theta \neq \frac{1}{2}$, the expression for Δx in (6.27) now becomes

$$\Delta x = k\Delta t a(x), \tag{6.29}$$

where

$$k = \sqrt{1 + 12\left(\theta - \tfrac{1}{2}\right)^2}. \tag{6.30}$$

We assume Δt is fixed and introduce a variable mesh along the x axis; so define $\Delta x_{i+1} := x_{i+1} - x_i$ for $i = 0, \ldots, I-1$. The spatial step length is defined from (6.29) and is taken to be the cell average since this is how the speed $a(x)$ is approximated in (6.25) (i.e. the term $\bar{A}$). Hence

$$\Delta x_{i+1} = k\Delta t a(x_{i+1/2}) =: \tfrac{1}{2}k\Delta t\left[a(x_{i+1}) + a(x_i)\right]. \tag{6.31}$$

In Figure 6-7 we demonstrate the effectiveness of using a variable spatial step length for the same example as in Figure 6-6 but when θ is the smaller root of (6.28). The oscillations and inaccuracies in the solution are eliminated in all cases.

We will not be able to eliminate the oscillations so effectively in 2D. The coefficients of the dispersion terms cannot be set to zero because both the spatial step lengths Δx and Δz will be functions of x and z. However, we can use the analysis from this section as a guide.

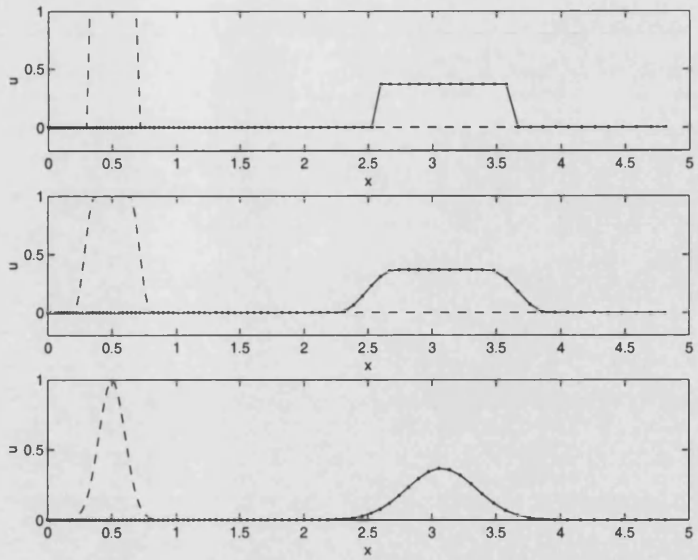

Figure 6-7: *Plots of the weighted box scheme ($\theta \approx 0.505$ from (6.28)) with variable $\triangle x$ (from (6.27)) applied to (6.24) with $a(x) = 1 + x$ and $\lambda = 1$ at time $t = 1$ for a square pulse (top plot), a sin-squared pulse (middle plot) and a Gaussian pulse (bottom plot). The dashed line denotes the initial condition, the dots joined by an unbroken line denotes the numerical solution and the thin unbroken line denotes the exact solution, with $\triangle t = 0.02$.*

6.4 The mine tailings problem in 2D

As discussed in the Introduction to this Chapter, Walter et al. (1994*a*) consider the mobility of potentially toxic dissolved metals discharged from a mine tailings source into an aquifer (see Figure 6-1). The flow is assumed to be incompressible and so we wish to solve (6.4) with the conditions on the boundaries given in Figure 6-8, i.e.

$$\nabla^2 \phi(x,z) = 0, \tag{6.32}$$

$$\phi_z(x,0) = 0, \quad \phi_z(x,Z) = -0.5, \quad \phi_x(0,z) = 0, \quad \phi(X,z) = 1. \tag{6.33}$$

The solution of this can easily be found using separation of variables. It is given by

$$\phi(x,z) = 1 + \sum_{n=1}^{\infty} \frac{X(-1)^n}{\pi^2 (n-1/2)^2 \sinh \frac{(n-1/2)\pi Z}{X}} \cosh \frac{(n-1/2)\pi z}{X} \cos \frac{(n-1/2)\pi x}{X}.$$

For simplicity we take $n = 1$. Then ϕ_x and ϕ_z are

$$\phi_x(x,z) = \frac{2}{\pi \sinh\left(\frac{\pi Z}{2X}\right)} \cosh\left(\frac{\pi z}{2X}\right) \sin\left(\frac{\pi x}{2X}\right) \tag{6.34}$$

$$\phi_z(x,z) = \frac{-2}{\pi \sinh\left(\frac{\pi Z}{2X}\right)} \sinh\left(\frac{\pi z}{2X}\right) \cos\left(\frac{\pi x}{2X}\right). \tag{6.35}$$

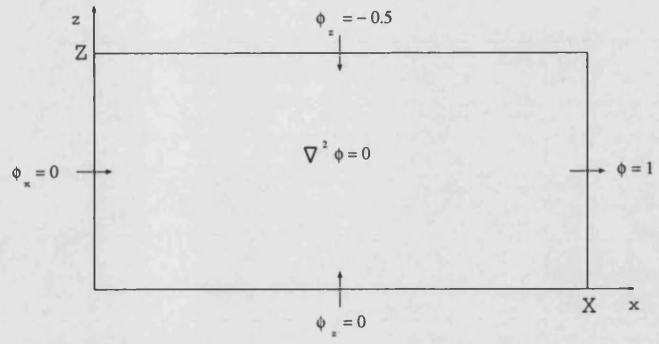

Figure 6-8: *2D cross-sectional plane showing the problem to solve for ϕ including the boundary conditions.*

Consider the system (6.5) in scalar form with a linear reaction term, i.e.

$$u_t + a(x,z)u_x + b(x,z)u_z = -\lambda u, \tag{6.36}$$

where $a \equiv \phi_x$, $b \equiv \phi_z$ and $\lambda > 0$. A modified equation analysis of the weighted box scheme applied to this problem is very complicated but it can be shown that, as for the 1D case, the U_{xx}, U_{zz}, U_{xxx} and U_{zzz} terms can be eliminated by setting

$$\triangle x^2 - \triangle t^2 a^2 - 12\left(\theta - \tfrac{1}{2}\right)^2 \triangle t^2 a^2 = 0 \tag{6.37}$$

$$\triangle z^2 - \triangle t^2 b^2 - 12\left(\theta - \tfrac{1}{2}\right)^2 \triangle t^2 b^2 = 0 \tag{6.38}$$

$$\left(\theta - \tfrac{1}{2}\right) - \lambda\left(\theta - \tfrac{1}{2}\right)^2 \triangle t - \tfrac{1}{4}\lambda \triangle t = 0. \tag{6.39}$$

However, since a and b depend on x and z, there are additional U_{xz}, U_{xxz} and U_{xzz} terms in the modified equation expansion whose coefficients cannot be set to zero. We still choose $\triangle x$ and $\triangle z$ to satisfy (6.37) and (6.38), with θ as the smaller root of (6.39), to investigate whether the oscillations are reduced. Hence set

$$\triangle x = ka(x,z)\triangle t, \qquad \triangle z = kb(x,z)\triangle t, \tag{6.40}$$

where k is defined in (6.30). The procedure carried out for the 1D case cannot be applied directly here because $\triangle x$ and $\triangle x$ now depend on both x and z. To overcome this problem we average a and b by integrating over z and x respectively, i.e. set

$$\hat{a}(x) = \frac{1}{Z}\int_0^Z a(x,z)\,\mathrm{d}z, \qquad \hat{b}(z) = \frac{1}{X}\int_0^X b(x,z)\,\mathrm{d}x. \tag{6.41}$$

These replace a and b in the expressions in (6.40) and so

$$\triangle x_{i+1} = \tfrac{1}{2}k\triangle t\left[\hat{a}(x_{i+1}) + \hat{a}(x_i)\right], \qquad \triangle z_{j+1} = \tfrac{1}{2}k\triangle t\left[\hat{b}(z_{j+1}) + \hat{b}(z_j)\right]. \tag{6.42}$$

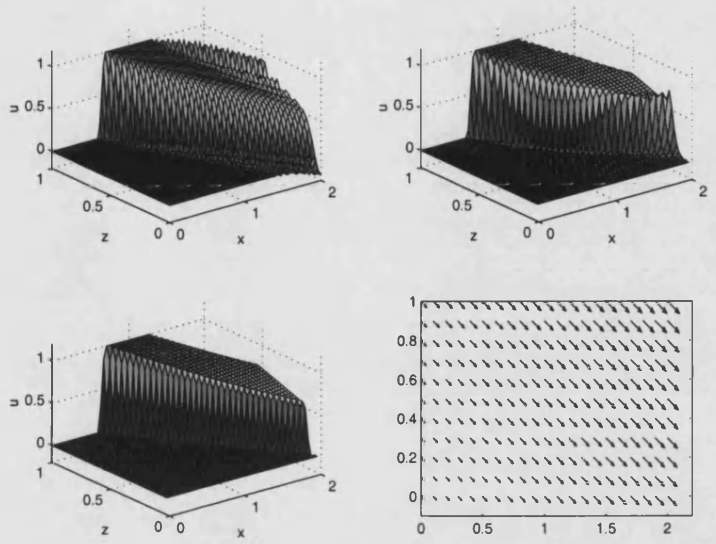

Figure 6-9: *The weighted box scheme applied to (6.36) with (6.43) and (6.44) where* $a_0 = a_1 = a_2 = 0.75$ *and* $b_0 = b_1 = b_2 = 0.5$, $\lambda = 1$ *and* $\triangle t = 0.02$. *The mesh is constant in the top left plot (with* $\triangle x = \triangle z = 0.025$*) and bottom left plot (using (6.47)) and variable in the top right plot (using (6.42)). The direction of the fluxes are shown in the bottom right plot..*

6.4.1 Linear fluxes

In this Section we suppose a and b are of the form

$$a(x,z) = a_0 + a_1 x + a_2 z \tag{6.43}$$

$$b(x,z) = -b_0 - b_1 x - b_2 z, \tag{6.44}$$

with $a_i,\ b_i \geq 0$ for $i = 0, 1, 2$. We consider this simple example because the exact solution of (6.36) with these linear fluxes can be found relatively easily. It is given by

$$u(x,z,t) = f(x_0, t-r)e^{-\lambda r}, \tag{6.45}$$

where again $f(x,t) = u(x,Z,t)$, and x_0 and r are found in terms of x and z by solving the following coupled pair of ODEs (with $x = x_0$ and $z = Z$ at $r = 0$):

$$\frac{\mathrm{d}x}{\mathrm{d}r} = a_0 + a_1 x + a_2 z, \quad \frac{\mathrm{d}z}{\mathrm{d}r} = -b_0 - b_1 x - b_2 z, \tag{6.46}$$

Figure 6-9 shows the weighted box scheme (with θ as the smaller root of (6.39)) applied to this problem. We have taken $\triangle t = 0.02$ and $\lambda = 1$ with boundary data given by (6.20) and (6.21) from Section 6.2.1 (where $T_s = 1$ and the results are plotted at $t = 3$).

	l^2 norm	maximum error
arbitrary constant mesh ($\Delta x = \Delta z = 0.025$)	0.31864	1.66288
variable mesh (6.42)	0.536114	0.18911
constant mesh (6.47)	0.08418	0.03887

Table 6.1: *Table showing a comparison of the l^2 norm and the maximum errors for the weighted box scheme with variable and constant meshes applied to (6.36) with (6.43) and (6.44) where $a_0 = a_1 = a_2 = 0.75$, and $b_0 = b_1 = b_2 = 0.5$.*

The velocity profiles are given in the bottom right figure. The top left plot shows the results of a constant mesh (chosen arbitrarily) with $\Delta x = \Delta z = 0.025$. There are severe oscillations which cannot be reduced if θ is increased. In the top right plots we have used the variable mesh given in (6.42). The results are much better although there are a couple of spikes near the $x = X$ boundary.

Instead of choosing an arbitrary constant mesh, we could also try to use the modified equation expansion to obtain a more appropriate mesh. To obtain formulas for Δx_i and Δz_j in (6.42) we averaged the expressions in (6.40) by integrating over z and x respectively. We could average in both the x and z directions, i.e.

$$\Delta x = \frac{1}{XZ}\int_0^Z \int_0^X a(x,z)\,\mathrm{d}x\,\mathrm{d}z, \qquad \Delta z = \frac{1}{XZ}\int_0^Z \int_0^X b(x,z)\,\mathrm{d}x\,\mathrm{d}z. \tag{6.47}$$

The bottom left plot in Figure 6-9 shows the results using a fixed mesh but with these choices of Δx and Δz. This is a great improvement on both the arbitrary constant and variable meshes (as shown in Table 6.1) and suggests that varying the mesh is not a good idea. However, we will now show that it is able to give more accurate results.

6.4.2 Separable (and related) fluxes

In Section 6.2.2 we briefly discussed an example with fluxes given by (6.43) and (6.44) where a_2 and b_1 were zero so the equation was separable. Hence we solved

$$u_t + (a_0 + a_1 x)u_x - (b_0 + b_2 z)u_z = -\lambda u. \tag{6.48}$$

If the mesh is now varied then the equations in (6.40) can be used directly (as for the 1D case) since a and b are only functions of x and z respectively. Figure 6-10 shows numerical results; the top plots use the weighted box scheme with a variable mesh (6.42) and the bottom plots use the constant mesh (6.47). The variable grid gives a much more accurate solution. This is to be expected since the problem can be converted into 1D where we know the variable mesh is more accurate.

We now increase a_2 and b_1 from zero and use the formulae in (6.42) to calculate Δx_i

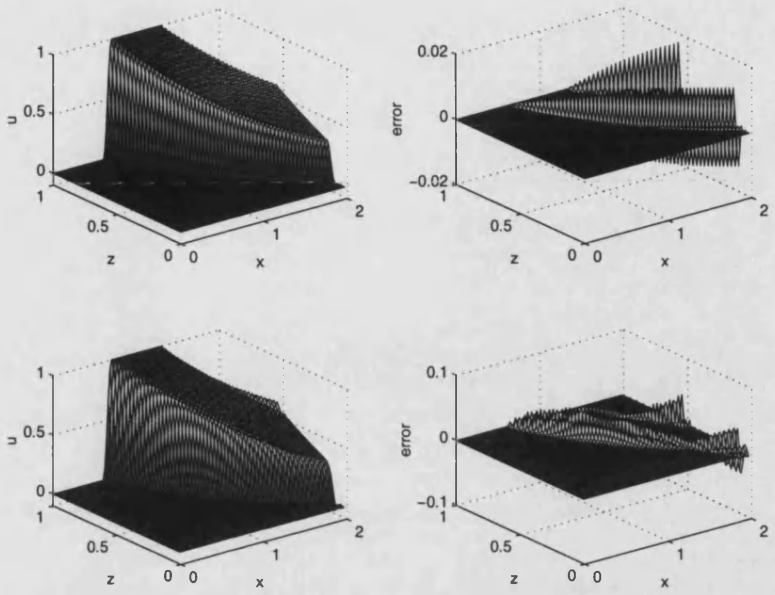

Figure 6-10: *The weighted box scheme applied to (6.36) with (6.43) and (6.44) where* $a_0 = a_1 = b_0 = b_2 = 0.5$, $a_2 = b_1 = 0$, $\lambda = 1$ *and* $\triangle t = 0.02$. *In the top plots the mesh is variable (using (6.42)) and in the bottom plots the mesh is constant (using (6.47)).*

and $\triangle z_j$. Table 6.2 shows the l^2 norm and the maximum error for various a_2 and b_1. The other parameters are all fixed. Varying the mesh gives better results up until $a_2 = b_1 = 0.2$. In Figure 6-11 we have plotted this case for both the variable and constant meshes. We see that, although the variable mesh gives a larger error, it has fewer oscillations (in the bottom left plot there are oscillations at the back of the pulse which are not there in top left plot). Unfortunately, as a_2 and b_1 increase this does not continue to be the case. We do not plot the results but more oscillations appear with the variable mesh whereas the oscillations for the constant mesh are reduced.

These results show that the variable mesh can give more accurate results provided a_2 and b_1 are not too large. However, this linear example is a special case because the velocity fluxes do not vary significantly throughout the domain. We can see this from observing the velocity fluxes in Figure 6-12; it shows four examples of a_i and b_i. The most pronounced variation occurs when the a and b are close to being separable (i.e. the top left example) and this is where the variable mesh is more accurate. In the top right and bottom left example the variation is small and we have seen that the constant mesh is better. Finally, in the bottom right example the variation is quite pronounced and in this case the variable mesh is much more accurate (shown in Figure 6-13). This is probably due to the fact that b_1 is small compared with b_0 and b_2 but it shows that the variable mesh can give a great improvement.

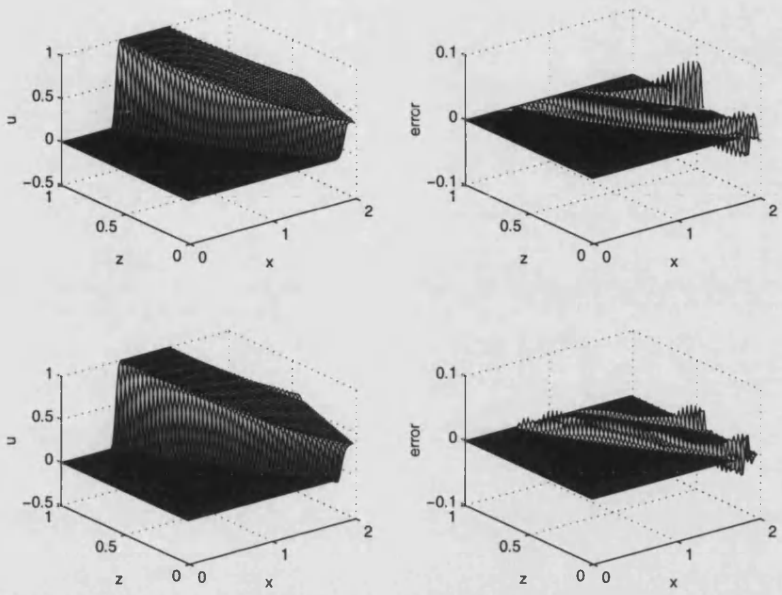

Figure 6-11: *The weighted box scheme applied to (6.36) with (6.43) and (6.44) where* $a_0 = a_1 = b_0 = b_2 = 0.5$, $a_2 = b_1 = 0.2$, $\lambda = 1$ *and* $\triangle t = 0.02$. *In the top plots the mesh is variable (using (6.42)) and in the bottom plots the mesh is constant (using (6.47)).*

	variable mesh (6.42)		**constant mesh (6.47)**	
	l^2 norm	maximum error	l^2 norm	maximum error
$a_2 = b_1 = 0$	0.05620	0.01765	0.22400	0.05476
$a_2 = b_1 = 0.01$	0.05867	0.01458	0.22366	0.05485
$a_2 = b_1 = 0.05$	0.06672	0.02537	0.21771	0.05438
$a_2 = b_1 = 0.1$	0.10064	0.03707	0.20430	0.05350
$a_2 = b_1 = 0.15$	0.13539	0.05898	0.18033	0.05146
$a_2 = b_1 = 0.18$	0.16410	0.05939	0.16735	0.05161
$a_2 = b_1 = 0.2$	0.17932	0.06615	0.15739	0.05030
$a_2 = b_1 = 0.3$	0.24046	0.09786	0.11359	0.03634
$a_2 = b_1 = 0.4$	0.29947	0.11599	0.07314	0.02569

Table 6.2: *Table showing a comparison of the weighted box scheme with variable and constant meshes applied to (6.36) with (6.43) and (6.44) where* $a_0 = a_1 = b_0 = b_2 = 0.5$ *and* a_2 *and* b_1 *are increased from zero.*

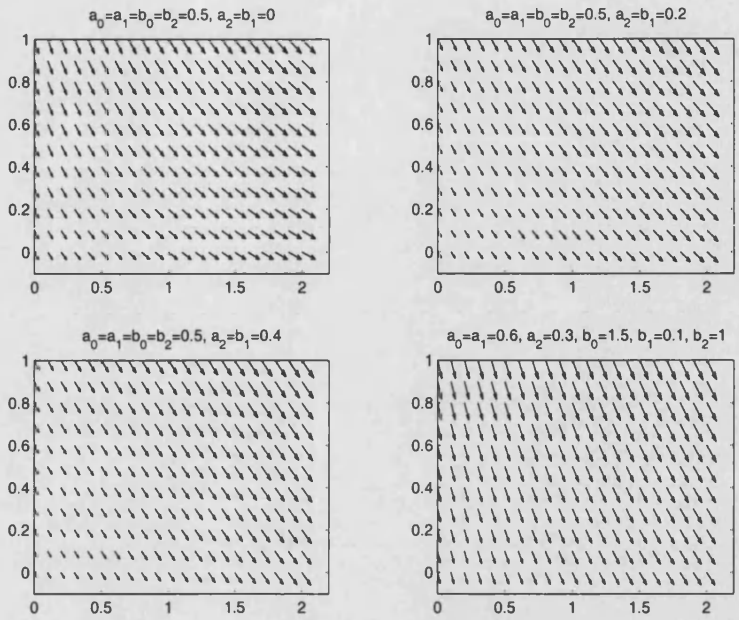

Figure 6-12: *Velocity profiles for four examples of the fluxes (6.43) and (6.44); top left:* $a_0 = a_1 = b_0 = b_2 = 0.5$ *and* $a_2 = b_1 = 0$ *(the separable case), top right:* $a_0 = a_1 = b_0 = b_2 = 0.5$ *and* $a_2 = b_1 = 0.2$, *bottom left:* $a_0 = a_1 = b_0 = b_2 = 0.5$ *and* $a_2 = b_1 = 0.4$, *bottom right:* $a_0 = a_1 = 0.6$, $a_2 = 0.3$, $b_0 = 1.5$, $b_1 = 0.1$ *and* $b_2 = 1$.

Hence, for a simple linear flux function, using a variable mesh can be advantageous over a constant mesh. However, this is not the case when there is little variation in the direction of the fluxes but this is to be expected and the constant mesh works well here. In practical situations the fluxes will be nonlinear and the velocity profiles will vary a great deal; the variable mesh should then be an improvement.

An extreme example

Firstly, we consider another example for the linear flux case which is the most different from the 1D situation. In the previous separable case, we took $a_2 = b_1 = 0$ and then increased these parameters from zero. This is similar to the 1D case. Suppose now that a is a function of z only (and/or b is a function of x only) and, for simplicity, assume $a_1 = b_1 = b_2 = 0$. Then we can really test how well the weighted box scheme works in 2D since it is the most extreme from the 1D situation. Hence we solve

$$u_t + (a_0 + a_2 z)u_x - b_0 u_z = -\lambda u. \tag{6.49}$$

To satisfy (6.37) and (6.38) choose

$$\triangle x = k(a_0 + a_2 z)\triangle t, \qquad \triangle z = k b_0 \triangle t. \tag{6.50}$$

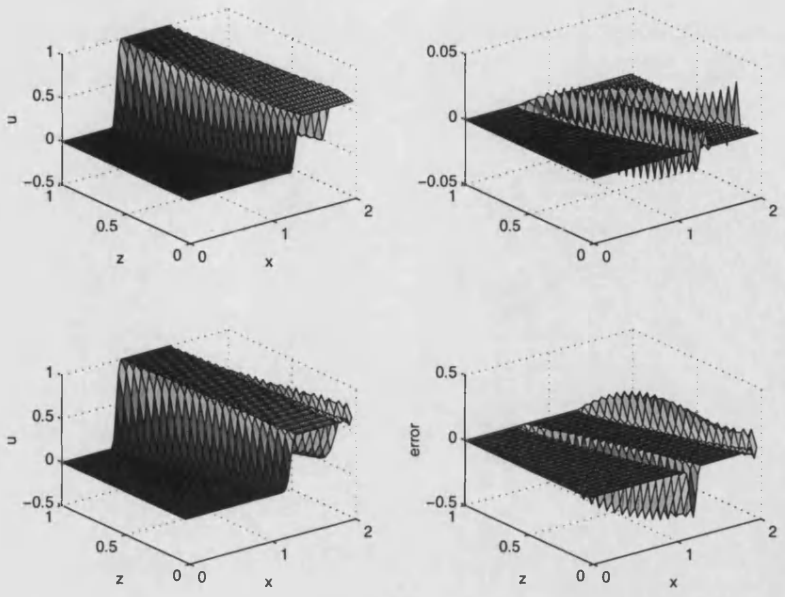

Figure 6-13: *The weighted box scheme applied to (6.36) with (6.43) and (6.44) where* $a_0 = a_1 = 0.6$, $a_2 = 0.3$, $b_0 = 1.5$, $b_1 = 0.1$ *and* $b_2 = 1$, $\lambda = 1$ *and* $\triangle t = 0.02$. *In the top plots the mesh is variable (using (6.42)) and in the bottom plots the mesh is constant (using (6.47)).*

Following the procedure above we have to integrate $\triangle x$ with respect to z, i.e.

$$\triangle x = \frac{k\triangle t}{Z}\int_0^Z (a_0 + a_2 z)\,\mathrm{d}z = k\triangle t(a_0 + \tfrac{1}{2}a_2 Z). \tag{6.51}$$

This will give a constant mesh in both directions. We could obtain a variable mesh in the x direction but which is constant between two values of z. For example, consider the left diagram in Figure 6-14. Suppose $U_{i,j}^n$ is known (for all i and j) and $U_{i,0}^{n+1}$, $U_{0,1}^{n+1}$ are also known. Then $U_{i,1}^{n+1}$ can be found with the following mesh:

$$\triangle x_1 = k\big(a_0 + \tfrac{1}{2}a_2(z_0 + z_1)\big)\triangle t, \qquad \triangle z = kb_0\triangle t. \tag{6.52}$$

However, to find $U_{i,2}^{n+1}$, a new mesh would have to be defined and, since the values $U_{i,1}^{n+1}$ are not at the same points (as is shown in Figure 6-14), an interpolation would have to be carried out. This is not a very satisfactory solution because it is computationally expensive and interpolating between the points will affect the accuracy.

In the right diagram Figure 6-14 the characteristics have been drawn for this situation. Instead of deriving the box scheme by integrating over a rectangular mesh (as discussed in Chapter 3), one option would be to integrate over a quadrilateral mesh with the differential equation in divergence form.

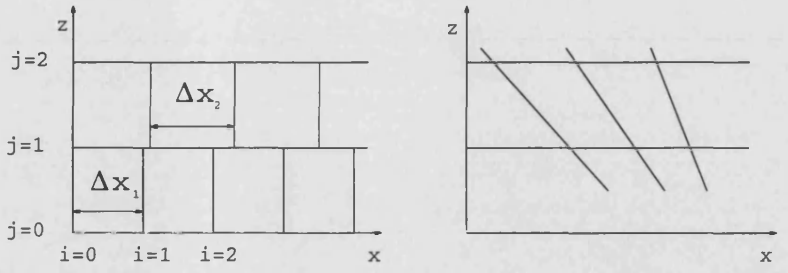

Figure 6-14: *The left figure shows an example of how we could create a mesh for the case when the weighted box scheme applied to (6.36) with (6.43) and (6.44) with* $a_1 = b_1 = b_2 = 0$. *The right figure shows the characteristics in this case.*

On the other hand, we can examine the constant mesh case for this extreme example with $\triangle x$ chosen from (6.51) and $\triangle z = kb_0\triangle t$. Figure 6-15 shows numerical results: in the top case $a_0 = b_0 = 1$ and $a_2 = 2$ and in the bottom case $a_0 = b_0 = 1$ and $a_2 = 0.5$. The l^2 norm and maximum error are shown in the first set of data in Table 6.3. Since the mesh is constant we observe the same phenomena as before: the constant mesh is more accurate when there is less variation in the direction of the fluxes (the top right plot has more variation then the bottom right).

We could reduce the oscillations by splitting the domain up into two (or more) sub-domains. Figure 6-15 shows that if the domain is split into two about the line $z = \frac{1}{2}Z$ then the fluxes are in roughly the same direction in each region. The mesh will still be constant in these regions but $\triangle x$ will be different ($\triangle z$ will remain the same). Hence, there are now two spatial step lengths in the x direction, denoted by $\triangle x_T$ and $\triangle x_B$

$$\triangle x_T = \frac{1}{(Z - \frac{1}{2}Z)} \int_{\frac{1}{2}Z}^{Z} a(x,z)\,\mathrm{d}z, \quad \triangle x_B = \frac{1}{(\frac{1}{2}Z - 0)} \int_{0}^{\frac{1}{2}Z} a(x,z)\,\mathrm{d}z. \tag{6.53}$$

To implement this method in practice we need to specify the numerical solution U for each level n at the $z = \frac{1}{2}Z$ boundary. Firstly, the solution is found in the top region since data is specified on the $z = Z$ boundary. Then, since the two meshes do not line up at $z = \frac{1}{2}Z$, we must linearly interpolate to obtain data on the grid points corresponding to the bottom mesh. Figure 6-16 shows the results of splitting up the domain in this way. Observe that the oscillations have reduced significantly, which is also confirmed in the bottom set of data in Table 6.3; the l^2 norm and maximum errors in the top and bottom regions are shown. These are significantly smaller and so splitting up the domain is a very effective way of reducing the oscillations.

	$a_0 = b_0 = 1,\ a_2 = 2$		$a_0 = b_0 = 1,\ a_2 = 0.5$	
	l^2 norm	max error	l^2 norm	max error
constant mesh (same in whole domain)	1.12903	0.29460	0.33360	0.08869
constant mesh (same in half domain)	0.49751 (T) 0.18725 (B)	0.19854 (T) 0.10575 (B)	0.10880 (T) 0.06211 (B)	0.03864 (T) 0.02824 (B)

Table 6.3: *Table showing a comparison of the l^2 norm and the maximum errors for the weighted box scheme applied to (6.36) with (6.43) and (6.44) in the extreme case $a_1 = b_1 = b_2 = 0$. The mesh is constant with $\triangle z = kb_0\triangle t$ and $\triangle x$ chosen by (6.51) in the first set of data and by (6.53) in the second ($T \equiv$ top, $B \equiv$ bottom).*

6.4.3 A more complicated flux function

Finally, we consider the hyperbolic equation in 2D given by (6.36) with the simplified mine tailings velocity fluxes (6.34) and (6.35). Figure 6-17 shows the direction of the velocity profiles in this case. There is a great variation between these arrows: at the top left boundary they are almost vertical and at the bottom left they are almost horizontal. Since $\phi_x(0,z) = \phi_z(x,0) = 0$ we have an extra complication near the origin: if the mesh is varied throughout the whole domain then very small spatial steps will be needed near the origin. For convenience, we will simply consider a smaller region

$$\big\{(x,z) \mid 0.2 \le x \le X,\ 0.2 \le z \le Z\big\}.$$

There is still great variation in the direction of the fluxes and so will test how well the variable mesh works for a more complicated example. Since the exact solution is not known we only examine the numerical solution to observe the oscillations. Note that θ is still the smaller root of (6.39) but we now use $\triangle t = 0.04$. Also, the boundary data is again a sin-squared pulse but is shifted to the left slightly.

In the method using the variable mesh, $\triangle x_i$ and $\triangle z_j$ are calculated using (6.42) where a and b are replaced by ϕ_x and $-\phi_z$, i.e.

$$\hat{a}(x) = \frac{1}{Z}\int_0^Z \phi_x(x,z)\,\mathrm{d}z, \quad \hat{b}(z) = -\frac{1}{X}\int_0^X \phi_z(x,z)\,\mathrm{d}x. \tag{6.54}$$

The constant mesh is then calculated using (6.47). Figure 6-18 shows the results using the variable mesh (top plot) and constant mesh (bottom plot). In this case we observe more oscillations in the variable mesh behind the pulse and more oscillations (or ripples) in the constant mesh at the front of the pulse. Since we cannot find the exact l^2 norm or maximum error, it is hard to deduce which mesh gives more accurate results. The boundary data is also quite extreme because it is not very smooth. Experiments found that a smoother pulse does eliminate the oscillations and it is hard to see any difference

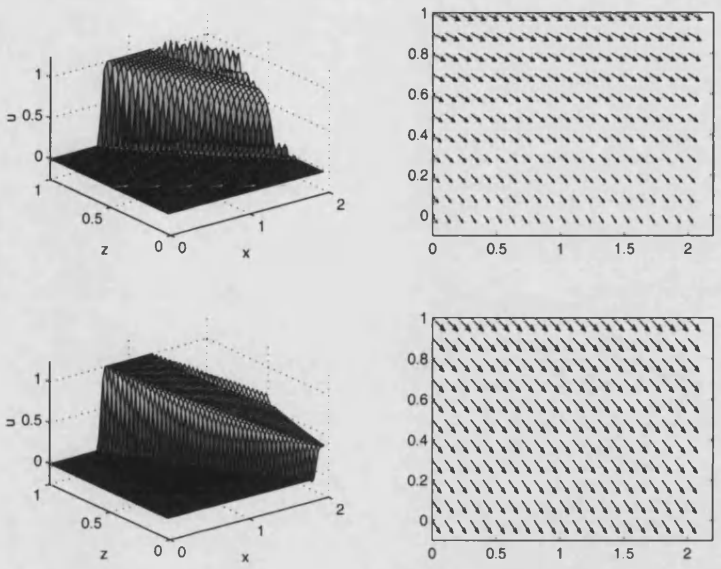

Figure 6-15: *The weighted box scheme applied to (6.36) with (6.43) and (6.44) for the extreme case where* $a_1 = b_1 = b_2 = 0$ *(*$\lambda = 1$ *and* $\triangle t = 0.02$*). In the top plots* $a_0 = b_0 = 1$, $a_2 = 2$ *and in the bottom plots* $a_0 = b_0 = 1$, $a_2 = 0.5$. *The mesh is constant with* $\triangle x$ *given by (6.51) and* $\triangle z = kb_0 \triangle t$.

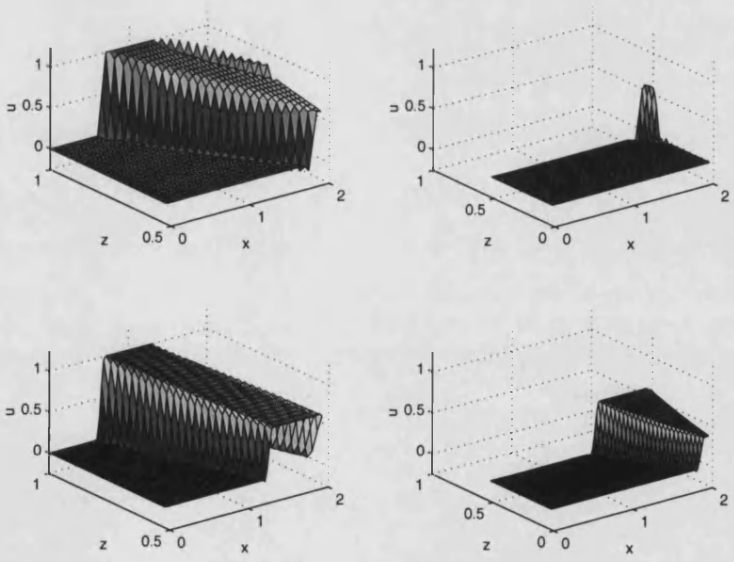

Figure 6-16: *The weighted box scheme applied to (6.36) with (6.43) and (6.44) for the extreme case where* $a_1 = b_1 = b_2 = 0$ *(*$\lambda = 1$ *and* $\triangle t = 0.02$*). In the top plots* $a_0 = b_0 = 1$, $a_2 = 2$ *and in the bottom* $a_0 = b_0 = 1$, $a_2 = 0.5$. *The domain is split into two regions with the top half shown on the left and the bottom half shown on the right.* $\triangle x$ *satisfies (6.53) and* $\triangle z = kb_0 \triangle t$.

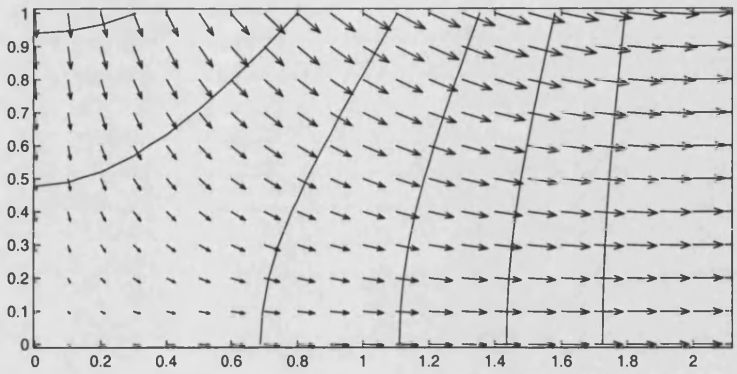

Figure 6-17: *Velocity fluxes for the simplified mine tailings example (6.34) and (6.35).*

between the variable and constant meshes in that case.

However, we would like to find a way to improve this solution. It makes sense to have a constant mesh in the right part of the domain as the bottom plot in Figure 6-18 does not have as many oscillations in this region. In a similar way to the extreme example considered above we split the domain into two sub-domains, but now along some line $x = \xi$. Suppose a variable mesh is used in the region $\{(x, z) \mid 0.2 \leq x \leq 1,\ 0.2 \leq z \leq 1\}$ and a constant mesh in the region $\{(x, z) \mid 1 \leq x \leq 2,\ 0.2 \leq z \leq 1\}$. The data on the $x = 1$ boundary between these two regions will have to be found using linear interpolation. The result of this modification can be seen in Figure 6-19: the left region is shown in the top plot and the right region in the bottom plot. We wish to compare this with the results when a variable mesh is used on the whole region (i.e. the top plot in Figure 6-18). This has been plotted again in Figure 6-20 with the regions split so an easy comparison can be made with Figure 6-19. Using a constant mesh in the right region does seem to eliminate the oscillations. A drawback is that more grid points have been used in the right region in Figure 6-19 than in Figure 6-20 but this seems to give a great improvement on using a variable mesh on the whole domain.

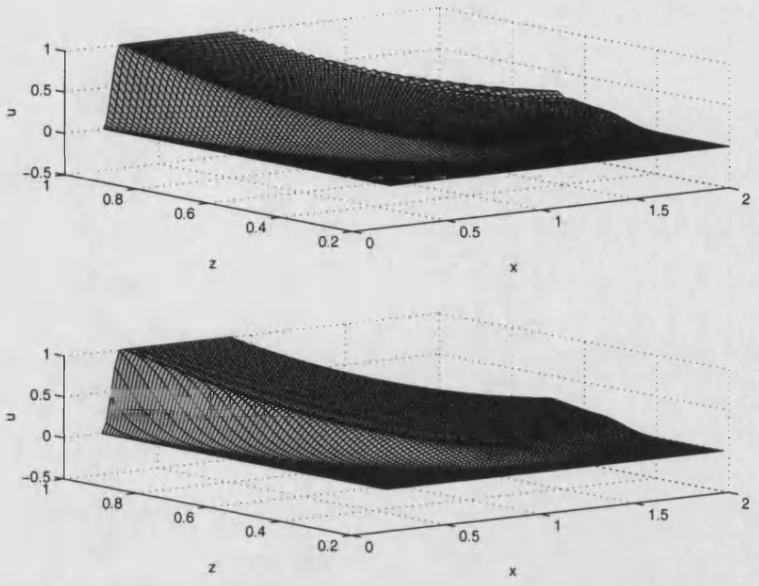

Figure 6-18: *The weighted box scheme applied to (6.36) with (6.34) and (6.35) and* $\triangle t = 0.02$. *In the top plot the mesh is variable (using (6.42)) and in the bottom plot the mesh is constant (using (6.47)). Note the domain is* $\{(x,z) \mid 0.2 \leq x \leq 2,\ 0.2 \leq z \leq 1\}$.

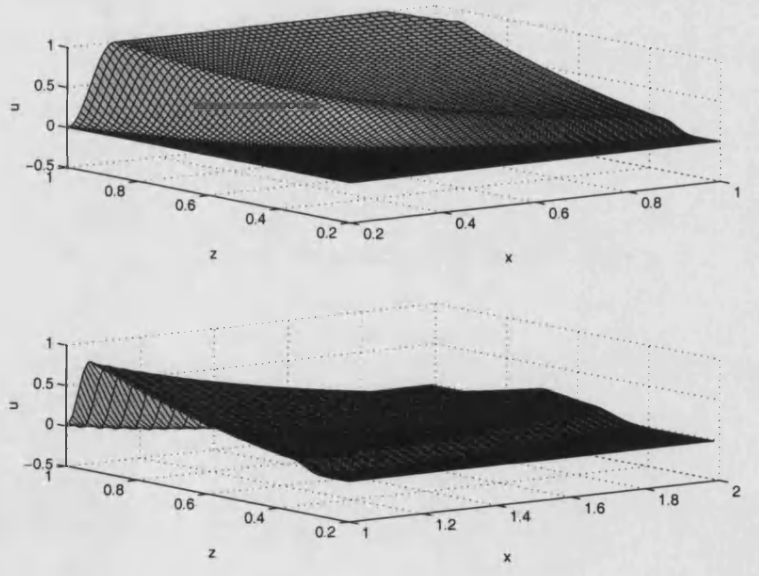

Figure 6-19: *The weighted box scheme applied to (6.36) with (6.34) and (6.35) and* $\triangle t = 0.02$. *The top plot shows the solution in the left half of the domain i.e.* $\{(x,z) \mid 0.2 \leq x \leq 1,\ 0.2 \leq z \leq 1\}$ *(with a variable mesh) and the bottom plot shows the solution in the right half of the domain i.e.* $\{(x,z) \mid 1 \leq x \leq 2,\ 0.2 \leq z \leq 1\}$ *(with a constant mesh).*

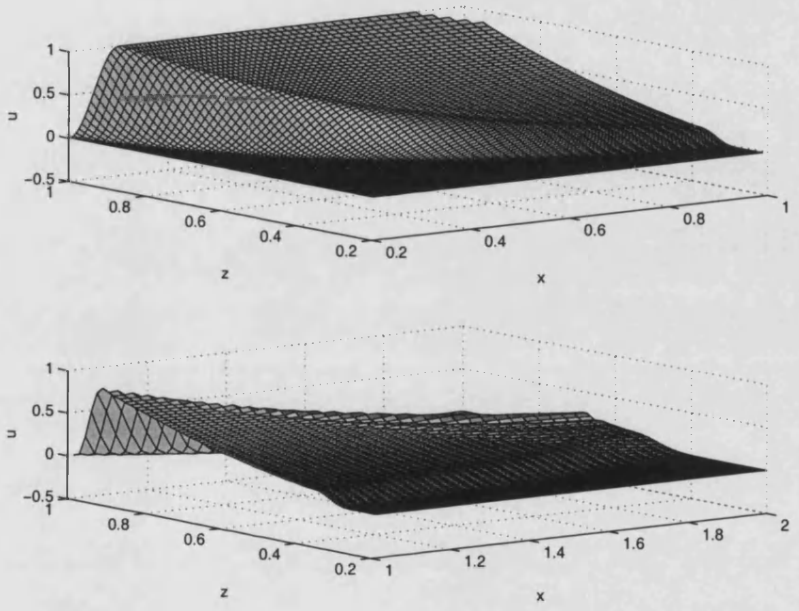

Figure 6-20: *The weighted box scheme applied to (6.36) with (6.34) and (6.35) and* $\triangle t = 0.02$. *The top plot shows the solution in the left half of the domain i.e.* $\{(x,z) \mid 0.2 \leq x \leq 1,\ 0.2 \leq z \leq 1\}$ *and the bottom plot shows the solution in the right half of the domain i.e.* $\{(x,z) \mid 1 \leq x \leq 2,\ 0.2 \leq z \leq 1\}$ *(both with variable meshes).*

Chapter 7

Conclusions and future work

In this work, through both extensive mathematical and numerical analysis, considerable insight has been gained into models describing the transport of chemical pollutants in groundwater flow. A key feature of these schemes is retardation, whereby the transport speed is much slower than the advection speed; this is not obvious from the original formulation of the models and so makes these systems both challenging and interesting to study.

We began in Chapter 2 by considering the mathematical analysis of the Linear Model and its extension to a general two equation model with a nonlinear source term. For the Linear Model, we discussed various analytical techniques which enabled us to predict the reduced speed, providing certain assumptions were made on the parameters λ and μ. Moreover, these techniques led us to show the role of diffusion in the problem. We also found, using analysis based on (Lighthill & Whitham 1955), that the first disturbance propagates with speed V but is damped exponentially and then the main disturbance lags behind and moves at the reduced speed. Guided by these results, we next considered the nonlinear source term and proved bounds on the speed of propagation for the nonlinear model, providing certain bounds existed on the partial derivatives of the reaction term. In proving this reduced speed result, we needed to show that the solution decayed exponentially as the chosen domain extended to infinity. We achieved this by combining a standard analysis of the theory of linear second order differential equations (whereby the model was integrated over the given domain) with a contraction mapping argument. This particular aspect of the two equation nonlinear model has not been found anywhere in the literature and is now complete; we have been able show how the phenomenon of reduced speed arises for a very general two equation model. However, it may be interesting and illuminating to prove these results for larger systems.

Our chosen numerical method is the *box scheme*, and in Chapter 3 this was applied to linear problems. There are numerous advantages of the box scheme: it is implicit,

unconditionally stable, second order in Δx and Δt and very compact (allowing the use of comparatively large time-steps). The main disadvantage of the box scheme is the presence of spurious oscillations in the numerical solution; these can be very bad for the linear advection equation with discontinuous data (because the chequerboard mode is not damped) but are also visible when the box-trap scheme is applied to the Linear Model (especially when the parameters are small and the boundary data is not smooth). These were found to be damped when a weighting was used in the time averaging of the spatial derivatives, although this does introduce diffusion into the scheme. However, this problem can effectively be ignored by ensuring the weighting parameter θ is chosen to be sufficiently close to $\frac{1}{2}$.

A modified equation analysis was then applied to our chosen numerical scheme which showed several interesting elements relating to how both the discretised equations and the differential model itself behaved. Firstly, the resulting expansion also gave the Improved-equilibrium model (found using the asymptotic analysis in Chapter 2) and so the theoretical phenomena of the reduced speed and diffusion could also be illustrated using this analysis. Secondly, it was able to predict where the observed oscillations would occur, found by separating the smooth and oscillatory parts of the numerical solution. Thirdly, by finding the expansion for the weighted box-trap scheme, it was able to give the best choice of the CFL number and the weighting parameter θ to reduce the oscillations. The modified equation analysis has not been used to this extent before and is shown to be a very useful tool in helping deal with the numerical difficulties encountered by the box scheme.

In Chapter 4 we successively adapted the box scheme to numerically solve nonlinear conservation laws with non-smooth data; which is well known to be very oscillatory around the discontinuity, and is typical of a second order method. We derived the scheme as a Petrov-Galerkin method to understand how these oscillations arose. This is an alternative interpretation of the scheme and enabled us to modify the trial space in the cell (or cells) containing the shock. The results of applying our algorithm showed how effectively the oscillations were eliminated. When shock-forming data was considered, small oscillations were still observed, but these were eliminated by weighting the scheme in the usual way. However, the algorithm was only tested for Burgers' equation and so it would be interesting to apply it to other examples. Also, for shock forming data, we had to predict where the shock formed as it made sense to use the box scheme without modification up until that point. Future work might involve experiments to obtain a more reliable way of predicting where the formation occurs.

The work discussed in the previous paragraph has applications to coupled reactive transport models; for example, we have shown that the solution of the Langmuir Model develops a steep front when the parameters λ and μ are large. We were again able to

reduce the oscillations that arise around the sudden switch in height by adapting the algorithm developed for nonlinear conservation laws. The results were very promising but this work could be extended by applying the algorithm to data that will develop into a steep front (instead of assuming one has already formed). This would also involve finding a criteria to decide at which point to start the algorithm.

In Chapter 5 we considered the Flushing-through Model; this is an extension of the Linear Model and a more realistic model problem. It incorporates a nonlinearity in the source term and potentially involves two speeds which adds complications to both the numerical and mathematical analysis previously applied to the Linear Model. The Improved-equilibrium model, which we derived for the total concentration d, was illustrated to only accurately describe the Flushing-through Model for large values of the parameters (and of similar order). This analysis does not seem to be a very useful tool in systems with more than two equations unless all the source terms are stiff. Its only real use is to show the presence of diffusion. However, the modified equation analysis of the weighted box-trap scheme applied to this model was again shown to be very informative in explaining the key features; namely the situation where the two species move at different speeds and that where all the concentrations moved at a speed other than the reduced or advected speed. In Chapter 3 we had only used this analysis on two equation linear systems: the work in this Chapter showed that useful deductions can be made for larger systems with nonlinear source terms.

We also discussed how the weighted box-trap scheme would be applied to a conservation law with a source term, which is now present in the Flushing-through Model. This difficulty had not previously been encountered in this Thesis since the Linear Model was written in such a way as to eliminate the source term from the Transport equation. Future work will involve an analysis to determine whether using a weighting to approximate the time derivative or simply the usual averaging gives a better numerical method.

Lastly in Chapter 5 we showed that the weighted box-trap scheme can be applied to systems with varying retardation speeds. We concluded in general that, when the parameters are small, the CFL number has to match the advection speed; but, as they increase, we can choose it to match the slowest propagation speed. However, we have only considered a very simple model problem but it is a good starting point as it exhibits the key features of larger systems. Future work could begin by applying the techniques discussed here to larger systems, for example the six equation model suggested by AEA Technology Harwell (now SERCO Assurance).

In Chapters 2-5 we considered 1D problems with constant speeds. In Chapter 6 we extended these models in two ways, firstly by assuming the fluxes are non-constant

and secondly by using two space dimensions. In practice it is much more realistic to analyse problems in 2D where the chemical pollutants would spread both vertically and horizontally through the water in time. Chapter 6 focused on applying the weighted box scheme to a simple hyperbolic equation with these extensions. To gain a better understanding of how the box scheme behaves for these more complicated problems, we began by considering a 1D equation with a non-constant velocity. Again a modified equation analysis proved key in reducing the oscillations; the expansion led to choosing a variable spatial step length which removed the dispersion term. These ideas were then applied to problems in 2D. We were able to show that, provided the direction of the velocity fluxes varied significantly within the domain, the variable mesh gave more accurate results and fewer oscillations. Also, we illustrated that a simple domain splitting approach was effective in improving the accuracy further in this situation (i.e. considerable change in the flux direction). If there was little variation in the direction of the velocity fluxes then a constant mesh seemed better, as long as the spatial step lengths were chosen by averaging to eliminate the dispersion terms in the modified equation expansion.

Further work for the nonlinear problem could involve developing a procedure to prevent the need for very small spatial steps near the origin (which was required in this case in order to satisfy the criteria for a variable mesh). A way to overcome this problem would be to use a constant mesh near the origin and so this idea could also be developed. Additionally, more work could be done on the domain splitting approach by investigating whether splitting the domain further would improve the accuracy of the numerical solution. Finally, our analysis of 2D problems has been confined to scalar examples and so this work should be continued by applying these findings to systems which more accurately describe reactive transport systems, such as the Linear or Flushing-through Models.

We are in the process of writing a paper on the mathematical and numerical analysis of the box scheme and its features with applications to reactive transport problems (Mitchell, Morton & Spence 2003*b*).

Appendix A

Extra results from Chapter 2

A.1 The Laplace transform solution

In Chapter 2 we stated the solution of the Linear Model as (2.22) with the function G defined by (2.23). If the boundary data is an injection of a short pulse of chemical pollutants into the groundwater, as defined in (2.24), then (2.22) becomes, for $t > \frac{x}{V}$ (the solution is zero for $t \leq \frac{x}{V}$)

$$a(x,t) = g\left(t - \tfrac{x}{V}\right) e^{-\frac{\lambda x}{V}} + \mu \int_{t-\frac{x}{V}-\delta}^{t-\frac{x}{V}} g\left(t - \tfrac{x}{V} - s\right) e^{-\frac{\lambda x}{V} - \mu s} G(s)\, \mathrm{d}s. \tag{A.1}$$

The first term gives the pulse moving down stream with speed V, but decaying exponentially and so is never observed in practice. Changing variables in (A.1) gives

$$a(x,t) = g\left(t - \tfrac{x}{V}\right) e^{-\frac{\lambda x}{V}} + \mu \int_0^{\delta} g(r) e^{-\frac{\lambda x}{V} - \mu\left(t - \frac{x}{V} - r\right)} G\left(t - \tfrac{x}{V} - r\right)\, \mathrm{d}r. \tag{A.2}$$

We now use the *Second Mean Value Theorem of the Integral Calculus*, see (Courant 1934, pages 256-257), which we can use to simplify the above expression. This was stated in Theorem 4 in Chapter 2 and is applied here to the integral in (A.2), where the interval $[t_1, t_2]$ is $[0, \delta]$. Hence, there exists $\tau \in [0, \delta]$ such that

$$\int_0^{\delta} g(r)\Gamma(r)\, \mathrm{d}r = \Gamma(0) \int_0^{\tau} g(r)\, \mathrm{d}r + \Gamma(\delta) \int_{\tau}^{\delta} g(r)\, \mathrm{d}r.$$

If we let $\tau \to \delta$ and use the normalising condition on the integral of g then the second integral disappears and so the second term in (A.2) becomes

$$a(x,t) = \alpha\mu e^{-\frac{\lambda x}{V} - \mu\left(t - \frac{x}{V}\right)} G\left(t - \tfrac{x}{V}\right), \tag{A.3}$$

since $\int_0^\tau g(r)\,\mathrm{d}r = \alpha$ and we have set

$$\Gamma(r) := \mu e^{-\frac{\lambda x}{V}-\mu\left(t-\frac{x}{V}-r\right)} G\left(t-\tfrac{x}{V}-r\right).$$

The expression in (A.3) can be written as

$$a(x,t) = \alpha\mu e^{-\frac{\lambda x}{V}-\mu\left(t-\frac{x}{V}\right)}\sqrt{\frac{\lambda x}{V\mu(t-x/V)}}\,I_1\left(2\sqrt{\tfrac{\lambda\mu x}{V}\left(t-\tfrac{x}{V}\right)}\right). \tag{A.4}$$

Now, when the modulus of the argument of a modified Bessel function (of integer order) is large we can use the following asymptotic expansion, as stated in (Abramowitz & Stegun 1965, page 377):

$$I_1(z) \sim \frac{e^z}{\sqrt{2\pi z}}\left\{1-\frac{3}{8z}-\frac{15}{2!(8z)^2}+\frac{315}{3!(8z)^3}-\ldots\right\}. \tag{A.5}$$

Hence, when λ and μ are large, the first term in (A.5) is sufficient and so

$$I_1\left(2\sqrt{\tfrac{\lambda\mu x}{V}\left(t-\tfrac{x}{V}\right)}\right) \sim \frac{1}{2\sqrt{\pi}}\exp\left(2\sqrt{\tfrac{\lambda\mu x}{V}\left(t-\tfrac{x}{V}\right)}\right)\left[\tfrac{\lambda\mu x}{V}\left(t-\tfrac{x}{V}\right)\right]^{-1/4}.$$

This can be substituted into (A.4) to give an approximation for a

$$a(x,t) \sim \frac{\alpha\mu}{2\sqrt{\pi}}\left[\frac{\lambda x/V}{(\mu(t-x/V))^3}\right]^{1/4}\exp\left[-\left(\sqrt{\mu\left(t-\tfrac{x}{V}\right)}-\sqrt{\tfrac{\lambda x}{V}}\right)^2\right], \tag{A.6}$$

which holds for $t > \frac{x}{V}$ and λ and μ are large. This expression has a maximum when

$$\sqrt{\mu\left(t-\tfrac{x}{V}\right)} = \sqrt{\tfrac{\lambda x}{V}},$$

which gives the value of t as stated in (2.26), i.e. $t = \frac{\lambda+\mu}{\mu V}x$. Thus the pulse moves at the reduced speed. Figure A-1 shows a three-dimensional diagram of how the profile of a in (A.6) behaves.

A.2 A correction to the Improved-equilibrium model

In Section 2.5 of Chapter 2 we derived the Equilibrium model and a first order correction (called the Improved-equilibrium model) by considering a general hyperbolic system with relaxation of the form (2.37). We now derive a second order correction to the Equilibrium model for the following linearised system (chosen to simplify the analysis):

$$\frac{\partial \mathbf{u}}{\partial t} + A\frac{\partial}{\partial x}\mathbf{u} + \frac{1}{\epsilon}\mathbf{S}(\mathbf{u}) = \mathbf{0}, \tag{A.7}$$

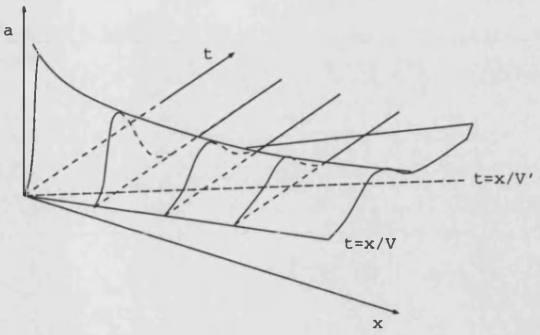

Figure A-1: *The exact Laplace transform solution of the Linear Model showing how the solution diffuses in time.*

and $\mathbf{S}$ is a linear function of $\mathbf{u}$. We determine whether the extra term found in this analysis will again match the modified equation expansion found in Chapter 3 when the box-trap scheme was applied (i.e. whether we obtain the first coefficient in the C_{xxx} term in (3.121)). The Improved-equilibrium model was derived by finding $\mathcal{M}^{(1)}[\mathbf{c}]$ which satisfied the following pair of equations:

$$\left(I - \frac{\partial}{\partial \mathbf{c}}(\mathcal{E}(\mathbf{c}))Q\right)\frac{\partial}{\partial x}\mathbf{f}(\mathcal{E}(\mathbf{c})) + \frac{\partial}{\partial \mathbf{u}}\mathbf{S}(\mathcal{E}(c))\mathcal{M}^{(1)}[\mathbf{c}] = \mathbf{0} \tag{A.8}$$

$$Q\mathcal{M}^{(1)}[\mathbf{c}] = \mathbf{0}. \tag{A.9}$$

This gave the ϵ term in the expansion of $\mathcal{M}^{\epsilon}[\mathbf{c}]$. However, we could also include the ϵ^2 term. After some analysis we find that $\mathcal{M}^{(2)}[\mathbf{c}]$ must satisfy

$$\left(I - \frac{\partial}{\partial \mathbf{c}}(\mathcal{E}(\mathbf{c}))Q\right)A\frac{\partial}{\partial x}\left(\mathcal{M}^{(1)}[\mathbf{c}]\right) - \frac{\partial}{\partial \mathbf{c}}\left(\mathcal{M}^{(1)}[\mathbf{c}]\right)QA\frac{\partial}{\partial x}(\mathcal{E}(\mathbf{c})) + \frac{\partial}{\partial \mathbf{u}}\mathbf{S}(\mathcal{E}(\mathbf{c}))\mathcal{M}^{(2)}[\mathbf{c}] = \mathbf{0}, \tag{A.10}$$

with $Q\mathcal{M}^{(2)}[\mathbf{c}] = \mathbf{0}$. Solving to find $\mathcal{M}^{(1)}[c]$ and $\mathcal{M}^{(2)}[c]$ leads to an improved correction to the local equilibrium approximation

$$\frac{\partial \mathbf{c}}{\partial t} + \frac{\partial}{\partial x}QA\left(\mathcal{E}(\mathbf{c}) + \epsilon\mathcal{M}^{(1)}[\mathbf{c}] + \epsilon^2\mathcal{M}^{(2)}[\mathbf{c}]\right) = \mathbf{0}. \tag{A.11}$$

We now apply this to the Linear Model.

A.2.1 The Linear Model

In Section 2.5.2 of Chapter 2 we found that

$$\mathcal{E}(c) = \begin{bmatrix} \frac{\mu}{\lambda+\mu}c \\ \frac{\lambda}{\lambda+\mu}c \end{bmatrix}, \qquad \mathcal{M}^{(1)}[c] = \begin{bmatrix} -\frac{V\lambda^2\mu}{(\lambda+\mu)^3}c_x \\ \frac{V\lambda^2\mu}{(\lambda+\mu)^3}c_x \end{bmatrix}.$$

These can now be substituted into (A.10) to find $\mathcal{M}^{(2)}[c]$ (also using $Q\mathcal{M}^{(2)}[\mathbf{c}] = \mathbf{0}$). After some manipulation this is found to be

$$\mathcal{M}^{(2)}[c] = \begin{bmatrix} \frac{V^2\lambda^3\mu(\lambda-\mu)}{(\lambda+\mu)^5}c_{xx} \\ -\frac{V^2\lambda^3\mu(\lambda-\mu)}{(\lambda+\mu)^5}c_{xx} \end{bmatrix}. \tag{A.12}$$

Hence a correction to the Improved-equilibrium model is found from (A.11), and so

$$c_t + \frac{V\mu}{\lambda+\mu}c_x - \frac{V^2\lambda\mu}{(\lambda+\mu)^3}c_{xx} + \frac{V^3\lambda\mu(\lambda-\mu)}{(\lambda+\mu)^5}c_{xxx} = 0. \tag{A.13}$$

As expected, the extra term gives the first term in the coefficient of C_{xxx} in the modified equation expansion for the box-trap scheme applied to the Linear Model (see equation (3.120) in Section 3.6 of Chapter 3 and the discussion in Section 3.6.1 afterwards).

A.3 Exact solution of the Linear Model using the domain of dependence

In Section 2 of Chapter 2 we stated that the exact solution (2.22) can be obtained using Laplace transforms. This result was merely quoted from (Rhee et al. 1986, page 159) as the details are very complicated. However, we now derive the result in a much cleaner way using Riemann's method applied to the second order equation. The problem is

$$a_{tt} + Va_{xt} = (\lambda+\mu)a_t - \mu V a_x, \tag{A.14}$$

$$a(x,0) = a_x(x,0) = a_t(x,0) = 0, \tag{A.15}$$

$$a(0,t) = g(t). \tag{A.16}$$

Consider the domain $D_V = OTR_VP$ as shown in Figure 2-4 in Chapter 2. Let P, T and R have co-ordinates $(0, t_P)$, $(x_Q, \frac{1}{V}x_Q)$ and $(x_Q, \frac{1}{V}x_Q + t_P)$ respectively. To solve (A.14) on D_V we first change variables to convert the problem into normal form. Define

$$y = \tfrac{1}{V}x, \qquad z = t - \tfrac{1}{V}x. \tag{A.17}$$

Then (A.14) reduces to

$$a_{yz} + \mu a_y + \lambda a_z = 0, \tag{A.18}$$

where the domain D_V is now $\Omega = OT'R'P'$ (see Figure 2-5 from Chapter 2). The points P', T' and R' have co-ordinates $(0, t_P)$, $(\frac{1}{V}x_Q, 0)$ and $(\frac{1}{V}x_Q, t_P)$ respectively.

The initial and boundary data are now

$$a(y,0) = a_y(y,0) = a_z(y,0) = 0, \tag{A.19}$$

$$a(0,z) = g(z). \tag{A.20}$$

We use Riemann's Method (see (Guenther & Lee 1988, pages 129-132) and (Garabedian 1964, pages 127-134)) which involves finding the solution in terms of definite integrals. Since λ and μ are constant, the equation (A.18) can be converted into canonical form. Setting

$$a(y,z) = u(y,z)e^{-\lambda y}e^{-\mu z}, \tag{A.21}$$

into (A.18) gives

$$u_{yz} - \lambda\mu u = 0. \tag{A.22}$$

Also, (A.19) and (A.20) become

$$u(y,0) = u_y(x,0) = u_z(x,0) = 0 \tag{A.23}$$

$$u(0,z) = e^{\mu z}g(z) =: h(z). \tag{A.24}$$

Define the linear differential operator $L[u]$ and the adjoint operator $M[v]$ by

$$L[u] = u_{yz} - \lambda\mu u, \qquad M[v] = v_{yz} - \lambda\mu v. \tag{A.25}$$

Then

$$vL[u] - uM[v] = (vu_z)_y - (uv_y)_z = \left(\tfrac{1}{2}vu_z - \tfrac{1}{2}v_z u\right)_y + \left(\tfrac{1}{2}u_y v - \tfrac{1}{2}uv_y\right)_z.$$

The above expression can be integrated over Ω and then the Divergence Theorem (2.126) applied to the right hand side. So

$$\begin{aligned}\iint_\Omega \left(vL[u] - uM[v]\right) \mathrm{d}y\,\mathrm{d}z &= \oint_{\partial\Omega} \left[\left(\tfrac{1}{2}u_z v - \tfrac{1}{2}uv_z\right)\mathrm{d}z - \left(\tfrac{1}{2}vu_y - \tfrac{1}{2}v_y u\right)\mathrm{d}y\right] \\ &= \oint_{\partial\Omega} \left[-uv_z + \tfrac{1}{2}(uv)_z\right]\mathrm{d}z - \left[-uv_y + \tfrac{1}{2}(uv)_y\right]\mathrm{d}y.\end{aligned} \tag{A.26}$$

Since $\mathrm{d}y = 0$ along OP' and $T'R'$, $\mathrm{d}z = 0$ along OT' and $P'R'$ and $u|_O = u|_{T'} = 0$ (from the first condition in (A.23)), (A.26) reduces to

$$\iint_\Omega \left(vL[u] - uM[v]\right) \mathrm{d}y\,\mathrm{d}z = (uv)_{R'} - (uv)_{P'} - \int_{T'}^{R'} uv_z\,\mathrm{d}z - \int_{P'}^{R'} uv_y\,\mathrm{d}y + \int_{O'}^{P'} uv_z\,\mathrm{d}z. \tag{A.27}$$

We now choose v to satisfy

$$M[v] = 0 \tag{A.28}$$

$$v_z = 0 \quad \text{on} \quad T'R' \tag{A.29}$$

$$v_y = 0 \quad \text{on} \quad P'R' \tag{A.30}$$

$$v = 1 \quad \text{at} \quad R'. \tag{A.31}$$

Then, substituting (A.28)–(A.31) into (A.27), along with $L[u] = 0$, gives

$$u|_{R'} = (uv)_{P'} - \int_O^{P'} uv_z \, \mathrm{d}z = \int_O^{P'} u_z v \, \mathrm{d}z, \tag{A.32}$$

since $uv_z = (uv)_z - u_z v$. Suppose the point R' has co-ordinates (ξ, η). Then (A.28)–(A.31) is known as the Goursat problem (Garabedian 1964, pages 117-119) for v, i.e.

$$v_{yz}(y,z) - \lambda\mu v(y,z) = 0, \quad y < \xi, \quad z < \eta \tag{A.33}$$

$$v_y(y,\eta) = 0, \quad y < \xi \tag{A.34}$$

$$v_z(\xi,z) = 0, \quad z < \eta \tag{A.35}$$

$$v(\xi,\eta) = 1. \tag{A.36}$$

We follow the method in (Garabedian 1964) which uses the method of successive approximations. Guenther & Lee (1988) also solve this problem, but assume the solution is of the form $v = \rho(r)$ where $r = (\xi - y)(\eta - z)$. Then the series solution of the resulting ODE is found for $\rho(r)$ which can be interpreted as a Bessel function.

In (Garabedian 1964, pages 118-120) the solution of a general second order equation in canonical form $u_{yz} = f(y, z, u, u_y, u_z)$, over a domain D (as shown in Figure 2-6 in Chapter 2 but with Ω replaced by D), is written in the form

$$u|_R = u|_P - u|_S + u|_Q + \iint_D f(y, z, u, u_y, u_z) \, \mathrm{d}y \, \mathrm{d}z. \tag{A.37}$$

A sequence of successive approximations is then defined in terms of this integral solution. We can do the same for (A.33)–(A.36), and so

$$v|_{R'} = v|_{P'} - v|_O + v|_{T'} + \lambda\mu \iint_\Omega v(y,z) \, \mathrm{d}y \, \mathrm{d}z. \tag{A.38}$$

We wish to find the solution on OP' which can be substituted into (A.32). To do this the solution within the region Ω must be found (since we already know the value of v at R'). We do not restrict the left hand corner of Ω to be at the origin, but instead label this S'. Then T' is the corresponding point horizontal to S' (and so need not be

on the y axis), and (A.38) can be rearranged to give

$$v|_{S'} = v|_{P'} - v|_{R'} + v|_{T'} + \lambda\mu \iint_{\Omega} v(y,z)\,\mathrm{d}y\,\mathrm{d}z. \tag{A.39}$$

From (A.34)–(A.36) it follows that $v|_{P'} = v|_{R'} = v|_{T'} = 1$; if S' has co-ordinates (α, β) then the sequence of successive approximations for (A.39) can be defined as

$$v_{n+1}(\alpha,\beta) = 1 + \lambda\mu \int_{\beta}^{\eta} \int_{\alpha}^{\xi} v_n(y,z)\,\mathrm{d}y\,\mathrm{d}z, \tag{A.40}$$

with $v_0(\alpha,\beta) = 1$. The solution of this is given in the following a Lemma.

Lemma 9. *The successive approximations v_n defined by (A.40) for the Goursat problem (A.33)–(A.36) are of the form*

$$v_n(\alpha,\beta) = \sum_{j=0}^{n} \frac{(\lambda\mu)^j(\xi-\alpha)^j(\eta-\beta)^j}{(j!)^2}. \tag{A.41}$$

Proof. We prove this by induction. The case $n = 0$ is obviously true since $v_0(\alpha,\beta) = 1$. Suppose the result is true for some $n > 0$. The right hand side of (A.40) is given by

$$\begin{aligned}
1 + \lambda\mu \int_{\beta}^{\eta} \int_{\alpha}^{\xi} v_n(y,z)\,\mathrm{d}y\,\mathrm{d}z &= 1 + \lambda\mu \left[\int_{\beta}^{\eta} \int_{\alpha}^{\xi} \left(\sum_{j=0}^{n} \frac{(\lambda\mu)^j(\xi-y)^j(\eta-z)^j}{(j!)^2} \right) \mathrm{d}y\,\mathrm{d}z \right] \\
&= 1 + \lambda\mu \sum_{j=0}^{n} \frac{(\lambda\mu)^j(\xi-\alpha)^{j+1}(\eta-\beta)^{j+1}}{[(j+1)!]^2} \\
&= 1 + \lambda\mu \sum_{k=1}^{n+1} \frac{(\lambda\mu)^{k-1}(\xi-\alpha)^k(\eta-\beta)^k}{(k!)^2} \\
&= 1 + \sum_{k=1}^{n+1} \frac{(\lambda\mu)^k(\xi-\alpha)^k(\eta-\beta)^k}{(k!)^2} \equiv v_{n+1}(\alpha,\beta).
\end{aligned}$$

Hence the result is true for $n+1$ which completes the proof. □

We now take the limit as $n \to \infty$ to find $v(\alpha,\beta)$. So

$$v(\alpha,\beta) = \lim_{n\to\infty} v_n(\alpha,\beta) = \sum_{j=0}^{\infty} \frac{\lambda\mu^j(\xi-\alpha)^j(\eta-\beta)^j}{(j!)^2}.$$

From (Abramowitz & Stegun 1965, page 375) this infinite sum can be expressed as a modified Bessel function of the first kind. Thus

$$v(\alpha,\beta) = I_0\big(2\sqrt{\lambda\mu(\xi-\alpha)(\eta-\beta)}\big). \tag{A.42}$$

Finally, let us consider (A.32). This can be written more explicitly as

$$u(\xi,\eta) = \int_0^\eta u_z(0,z)v(0,z)\,\mathrm{d}z. \tag{A.43}$$

Using (A.24) we have $u_z(0,z) = h'(z) = e^{\mu z}\left[\mu g(z) + g'(z)\right]$. Also, from (A.42), $v(0,z) = I_0\big(2\sqrt{\lambda\mu\xi(\eta - z)}\big)$. Hence the solution (A.43) becomes

$$u(\xi,\eta) = \int_0^\eta e^{\mu z}\left[\mu g(z) + g'(z)\right] I_0\big(2\sqrt{\lambda\mu\xi(\eta - z)}\big)\,\mathrm{d}z. \tag{A.44}$$

If, for convenience, we interchange the co-ordinates (ξ,η) and (y,z) then

$$u(y,z) = \int_0^z e^{\mu\eta}\left[\mu g(z) + g'(z)\right] I_0\big(2\sqrt{\lambda\mu y(z - \eta)}\big)\,\mathrm{d}\eta. \tag{A.45}$$

We can now use the conversion (A.21) to obtain an integral expression for a, i.e.

$$a(y,z) = e^{-\lambda y}\int_0^z e^{-\mu(z-\eta)}\left[\mu g(z) + g'(z)\right] I_0\Big(2\sqrt{\lambda\mu y(z - \eta)}\Big)\,\mathrm{d}\eta. \tag{A.46}$$

Finally, using (A.17) to convert back to (x,t), leads to the solution

$$a(x,t) = e^{-\frac{\lambda}{V}x}\int_0^{t-\frac{x}{V}} e^{-\mu\left(t-\frac{1}{V}x-\eta\right)}\left[\mu g(z)) + g'(z)\right] I_0\left(2\sqrt{\tfrac{\lambda\mu x}{V}\left(t - \tfrac{1}{V}x - \eta\right)}\right)\,\mathrm{d}\eta. \tag{A.47}$$

for $t > \frac{x}{V}$.

We can show that (A.47) is identical to the Laplace Transform solution (2.22) from Chapter 2. Consider (2.22), with $t > \frac{x}{V}$, i.e.

$$a(x,t) = g\left(t - \tfrac{x}{V}\right)e^{-\frac{\lambda x}{V}} + \mu e^{-\frac{\lambda x}{V}}\int_0^{t-\frac{x}{V}} g\left(t - \tfrac{x}{V} - s\right)e^{-\mu s}\sqrt{\tfrac{\lambda x}{\mu V s}}I_1\left(2\sqrt{\tfrac{\lambda\mu x s}{V}}\right)\,\mathrm{d}s. \tag{A.48}$$

Using the fact that

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[\frac{1}{\mu}I_0\left(2\sqrt{\tfrac{\lambda\mu x s}{V}}\right)\right] = \sqrt{\tfrac{\lambda x}{\mu V s}}I_1\left(2\sqrt{\tfrac{\lambda\mu x s}{V}}\right),$$

we can integrate (A.48) by parts to obtain

$$\begin{aligned} a(x,t) \;=\; & g\left(t - \tfrac{x}{V}\right)e^{-\frac{\lambda x}{V}} + \mu e^{-\frac{\lambda x}{V}}\left[g\left(t - \tfrac{x}{V} - s\right)e^{-\mu s}\tfrac{1}{\mu}I_0\left(2\sqrt{\tfrac{\lambda\mu x s}{V}}\right)\right]_0^{t-\frac{x}{V}} \\ & -\mu e^{-\frac{\lambda x}{V}}\int_0^{t-\frac{x}{V}}\left[-g'\left(t - \tfrac{x}{V} - s\right) - \mu g\left(t - \tfrac{x}{V} - s\right)\right]e^{-\mu s}\tfrac{1}{\mu}I_0\left(2\sqrt{\tfrac{\lambda\mu x s}{V}}\right)\,\mathrm{d}s. \end{aligned}$$

The first line of the above expression is zero, provided $g(0) = 0$. Hence, changing variables in the integral by setting $s = t - \frac{1}{V}x - \eta$ leads to (A.47).

A.4 Exponential decay of a simple ODE

In Section 2.7 of Chapter 2 we proved that the solution of the two equation model with general source term decayed exponentially to zero as the domain extended to infinity. As motivation for the techniques used in the proof we proved a similar result for a simple ODE. This is outlined here because it shows that proving a solution is exponentially decaying is much less trivial than proving it is bounded. Also, the reader is referred to this Appendix to help understand some of the steps in Chapter 2. We formulate the result in the following Lemma.

Lemma 10. *Consider*

$$y'(x) = \lambda(x)y(x), \qquad y(\xi) = 1, \tag{A.49}$$

where

$$\lambda(x) \leq -\lambda_0, \qquad \lambda_0 > 0, \tag{A.50}$$

and $|\lambda(x)|$ *is uniformly bounded for all* $x \in [\xi, +\infty)$. *Then*

$$|y(x)|e^{\lambda_0(x-\xi)} \leq K, \tag{A.51}$$

for all x, *where* K *is a positive constant, i.e.* y *decays at least as fast as* $e^{-\lambda_0(x-\xi)}$.

Proof. Set

$$z(x) = y(x)e^{(\lambda_0+\gamma)(x-\xi)}, \tag{A.52}$$

where $\gamma > 0$ is constant. We prove that z is bounded in an appropriately defined norm. It can easily be shown that z satisfies

$$z'(x) = \big[\lambda(x) + \lambda_0 + \gamma\big]z(x), \qquad z(\xi) = 1.$$

Define the operator T by

$$(Tz)(x) = 1 + \int_\xi^x \big[\lambda(t) + \lambda_0 + \gamma\big]z(t)\,\mathrm{d}t. \tag{A.53}$$

So

$$(Tz)(x) - (Tu)(x) = \int_\xi^x \big[\lambda(t) + \lambda_0 + \gamma\big](z(t) - u(t))\,\mathrm{d}t. \tag{A.54}$$

Since $|\lambda|$ is bounded we can assume the following:

$$0 \leq \lambda + \lambda_0 + \gamma \leq \frac{1}{n}\gamma, \tag{A.55}$$

for some constant $n > 1$. Also define

$$||z||_e = \max\{|z(x)|e^{-\alpha(x-\xi)}\}, \tag{A.56}$$

where $\alpha > 0$ is constant. Then, following the proof of existence and uniqueness of an initial value problem which is given in (Walter 1998, pages 62-64), we have

$$(Tz)(x) - (Tu)(x) = \int_\xi^x [\lambda(t) + \lambda_0 + \gamma](z(t) - u(t))e^{-\alpha(t-\xi)}e^{\alpha(t-\xi)}\,\mathrm{d}t,$$

and so

$$\begin{aligned} |(Tz)(x) - (Tu)(x)| &\leq \frac{1}{n}\gamma\int_\xi^x |z(t) - u(t)|e^{-\alpha(t-\xi)}e^{\alpha(t-\xi)}\,\mathrm{d}t \\ &\leq \frac{1}{n\alpha}\gamma||z-u||_e\left(e^{\alpha(x-\xi)} - 1\right) \\ &\leq \frac{1}{n\alpha}\gamma||z-u||_e e^{\alpha(x-\xi)}. \end{aligned}$$

Hence

$$||Tz - Tu||_e \leq \frac{1}{n\alpha}\gamma||z-u||_e. \tag{A.57}$$

This means that, provided

$$\frac{1}{n\alpha}\gamma < 1, \tag{A.58}$$

T is a contraction and so z is bounded in the norm (A.56), i.e.

$$|z(x)|e^{-\alpha(x-\xi)} \leq K,$$

for some positive K. Substituting in y from (A.52) leads to

$$|y(x)|e^{(\lambda_0+\gamma-\alpha)(x-\xi)} \leq K. \tag{A.59}$$

Hence the required result (A.51) holds if $\gamma - \alpha \geq 0$. Combining this condition with (A.58) gives

$$\frac{1}{n}\gamma < \alpha \leq \gamma. \tag{A.60}$$

The proof is therefore complete, provided (A.60) holds with $n > 1$. □

Appendix B

Extra results from Chapter 3

B.1 Exact solution of the discretisation of the box scheme

In this section we use a technique outlined by (Spiegel 1971, page 186) to solve the discretised equations for the box scheme when applied to a simple linear advection equation. This was briefly discussed in Section 3.2.1 of Chapter 3 but we now study the technique in more detail. To simplify the analysis, t_1 and t_2 are chosen in (3.23) so there is one non-zero value on the t-axis. For convenience we assume this occurs at $n = 0$. Then (3.23) becomes

$$U_0^n = \begin{cases} 1, & n = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{B.1}$$

We then need to effectively shift the x-axis by one node and so the initial condition in (3.23) is replaced by

$$U_j^{-1} = 0. \tag{B.2}$$

Figure B-1 shows the mesh for the box scheme with the data prescribed along the $x = 0$ and $t = -t^1$ $(= -\triangle t)$ axes. The diagonal line is the characteristic emanating from the origin and corresponds to setting $p = 1$ in (3.22) (and so $\lambda = 0$). As described in Chapter 3, the numerical solution can be written as

$$U_j^n = \left(\frac{\lambda + E_2^{-1}}{1 + \lambda E_2^{-1}} \right)^j C_n, \tag{B.3}$$

where the C_n, for $n \geq 0$, are found using the initial and boundary conditions and the operator E_2 is defined in (3.24). Assuming (B.1) and (B.2) hold, it is easy to see that

$$C_n = \begin{cases} 1, & n = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{B.4}$$

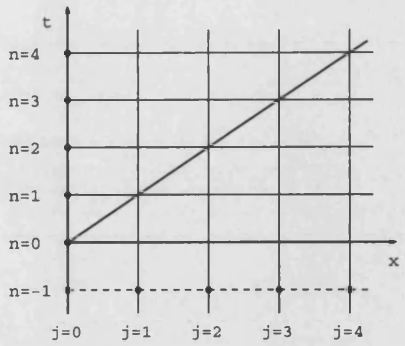

Figure B-1: *The mesh for the box scheme applied to the linear advection equation.*

Now, (B.3) can be written as

$$U_j^n = (\lambda + E_2^{-1})^j (1 + \lambda E_2^{-1})^{-j} C_n = \lambda^j \left(1 + \frac{E_2^{-1}}{\lambda}\right)^j (1 + \lambda E_2^{-1})^{-j} C_n.$$

We can expand both terms in the above expression and so

$$\begin{aligned} U_j^n &= \lambda^j \left(1 + \frac{E_2^{-1}}{\lambda}\right)^j \left[C_n - j\lambda C_{n-1} + \frac{j(j+1)}{2!}\lambda^2 C_{n-2} - \frac{j(j+1)(j+2)}{3!}\lambda^3 C_{n-3} + \ldots\right] \\ &= \lambda^j \left(\sum_{k=0}^{j} \left(\frac{j}{k}\right) \frac{E_2^{-k}}{\lambda^k}\right) \left[C_n - j\lambda C_{n-1} + \ldots + \frac{j(j+1)\ldots(j+n-1)}{n!}(-\lambda)^n C_0\right], \end{aligned} \tag{B.5}$$

where we have used (B.4) to eliminate any C_n's for $n < 0$. Using this formula expressions can be derived for fixed j and n in turn and we can then investigate what happens as the indices n and j tend to infinity respectively.

Let us first fix j and consider a general n. We can obtain a formula for U_j^n but this is only valid if n starts from the diagonal. We demonstrate this by writing down the first few cases. We already know that U_0^n is given by (B.1). Simple calculations give

$$\begin{aligned} U_1^n &= (-\lambda)^{n-1}(1-\lambda^2), && n = 1, 2, \ldots \\ U_2^n &= (-\lambda)^{n-2}(1-\lambda^2)\Big[(n-1) - (n+1)\lambda^2\Big], && n = 2, 3, \ldots \\ U_3^n &= \frac{(-\lambda)^{n-3}}{2!}(1-\lambda^2)\Big[(n-1)(n-2) - 2(n-1)(n+1)\lambda^2 + (n+1)(n+2)\lambda^4\Big], \\ & && n = 3, 4, \ldots \\ U_4^n &= \frac{(-\lambda)^{n-4}}{3!}(1-\lambda^2)\Big[(n-1)(n-2)(n-3) - 3(n-1)(n-2)(n+1)\lambda^2 \\ & \quad +3(n-1)(n+1)(n+2)\lambda^4 - (n+1)(n+2)(n+3)\lambda^6\Big], && n = 4, 5, \ldots, \end{aligned} \tag{B.6}$$

and so on. Suppose we now fix n and consider a general j. We can obtain similar expressions but, in this direction, they hold for all $j \geq 0$. So

$$\begin{aligned}
U_j^0 &= \lambda^j, \\
U_j^1 &= j\lambda^{j-1}(1-\lambda^2), \\
U_j^2 &= \frac{j\lambda^{j-2}}{2!}(1-\lambda^2)\Big[(j-1)-(j+1)\lambda^2\Big], \\
U_j^3 &= \frac{j\lambda^{j-3}}{3!}(1-\lambda^2)\Big[(j-1)(j-2)-2(j-1)(j+1)\lambda^2+(j+1)(j+2)\lambda^4\Big], \\
U_j^4 &= \frac{j\lambda^{j-4}}{4!}(1-\lambda^2)\Big[(j-1)(j-2)(j-3)-3(j-1)(j-2)(j+1)\lambda^2 \\
&\quad +3(j-1)(j+1)(j+2)\lambda^4-(j+1)(j+2)(j+3)\lambda^6\Big], \qquad \text{(B.7)}
\end{aligned}$$

and so on. From this we can observe a general pattern. In the former case (i.e. fixing j and letting n grow arbitrarily large) we obtain

$$\begin{aligned}
U_j^n = \frac{(-\lambda)^{n-j}}{(j-1)!}(1-\lambda^2)\Big[&(n-1)(n-2)\dots(n-j+1) \\
&-(j-1)(n-1)\dots(n-j+2)(n+1)\lambda^2 \\
&+\frac{(j-1)(j-2)}{2!}(n-1)\dots(n-j+3)(n+1)(n+2)\lambda^4 \\
&-\dots \\
&+(-1)^{j-1}(n+1)(n+2)\dots(n+j-1)\lambda^{2(j-1)}\Big], \qquad \text{(B.8)}
\end{aligned}$$

for $n \geq j$. In the latter case (i.e. fixing n and letting j grow arbitrarily large) we obtain a similar expression, which holds for all $j \geq 0$

$$\begin{aligned}
U_j^n = \frac{j\lambda^{j-n}}{n!}(1-\lambda^2)\Big[&(j-1)(j-2)\dots(j-n+1) \\
&-(n-1)(j-1)\dots(j-n+2)(j+1)\lambda^2 \\
&+\frac{(n-1)(n-2)}{2!}(j-1)\dots(j-n+3)(j+1)(j+2)\lambda^4 \\
&-\dots \\
&+(-1)^{n-1}(j+1)(j+2)\dots(j+n-1)\lambda^{2(n-1)}\Big], \qquad \text{(B.9)}
\end{aligned}$$

We now plot (B.8) and (B.9) for $j = 4$ and $n = 4$ respectively and consider what happens when n and j tend to infinity for various values of λ (with $|\lambda| < 1$). We consider λ positive and negative separately since $\lambda < 0$ corresponds to $p < 1$ and $\lambda > 0$ corresponds to $p > 1$. Figure B-2 shows the solution U_4^n against n as $n \to \infty$ from the diagonal (i.e. for $n \geq j$) for $\lambda = \pm 0.25, \pm 0.5, \pm 0.75$. We observe that there are only oscillations for $\lambda > 0$. These become more severe as $\lambda \to 1$. Also, as $\lambda \to 0^{\pm}$ the solution very quickly decays to zero which can easily be seen from (B.8). This is to

be expected since $\lambda = 0$ corresponds to $p = 1$ and we know there are no oscillations in this case. In Figure B-3 we have considered a more extreme situation: $\lambda = \pm 0.99$. Neither case is very likely to occur in practice as these correspond to $p = 199$ and $p = \frac{1}{199}$ respectively. However, this does illustrate that when p becomes very small (in the bottom plot) the solution very quickly tends to zero without oscillating.

Figures B-4 and B-5 show the solution U_j^4 against j as $j \to \infty$ starting from $j = 0$. We consider the same values of λ as in Figures B-2 and B-3. There are now oscillations if λ is negative, although now the solution does change sign when λ is positive. We can see this from examining U_j^4 in (B.7). The term in square brackets is a cubic for λ^2 and this corresponds to the number of zeros observed in Figure B-4 for both $\lambda > 0$ and $\lambda < 0$ (though for $\lambda < 0$ there are oscillations in between). In Figure B-5 we have again considered $\lambda = \pm 0.99$ and the same phenomenon is observed: for all values of λ the solution U_j^4 increases to a peak after going through the last root and then tends to zero. However, this convergence is much slower than for U_4^n.

This analysis, although not rigorous, has highlighted a of the feature of the box scheme which we observed in Chapter 3. The solution is translated along the diagonal at the mesh speed (independent of the wave speed) with oscillations of polynomial size trailing out either side. These vary in number depending on the size and sign of λ but will die away as n and j get large. However, the technique is very technical, even for the box scheme applied to the linear advection equation. The expression (B.5) is so complicated because the discretisation in (3.22) has two points at the new time level.

B.2 The ETIR Method

Consider the Linear Model written in the form

$$a_t + b_t + V a_x = 0 \tag{B.10}$$

$$b_t = -\lambda a + \mu b. \tag{B.11}$$

In Chapter 3 we briefly mentioned the ETIR method: this uses an explicit scheme to approximate (B.10) and an implicit scheme to approximate (B.11). Then the resulting discretised equations are given by

$$\left(A_j^{n+1} - A_j^n\right) + \left(B_j^{n+1} - B_j^n\right) + p\left(A_j^n - A_{j-1}^n\right) = 0 \tag{B.12}$$

$$B_j^{n+1} - B_j^n = \lambda' A_j^{n+1} - \mu' B_j^{n+1}, \tag{B.13}$$

where $p = V\Delta t/\Delta x$, $\lambda' = \lambda \Delta t$ and $\mu' = \mu \Delta t$. We can perform a comprehensive Fourier analysis to give a necessary condition for this method to be Lax-Richtymer stable.

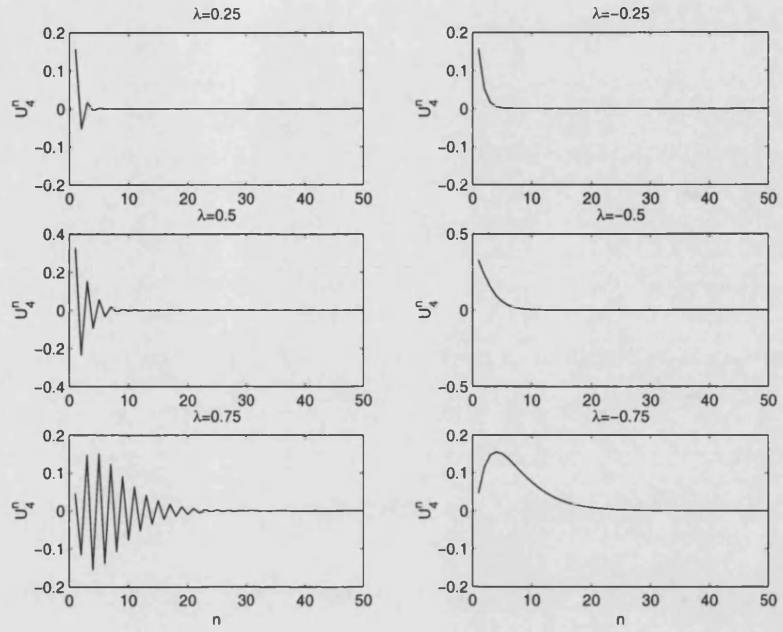

Figure B-2: U_4^n *for various values of* λ. *In the left three plots* $\lambda > 0$ *and in the right three plots* $\lambda < 0$.

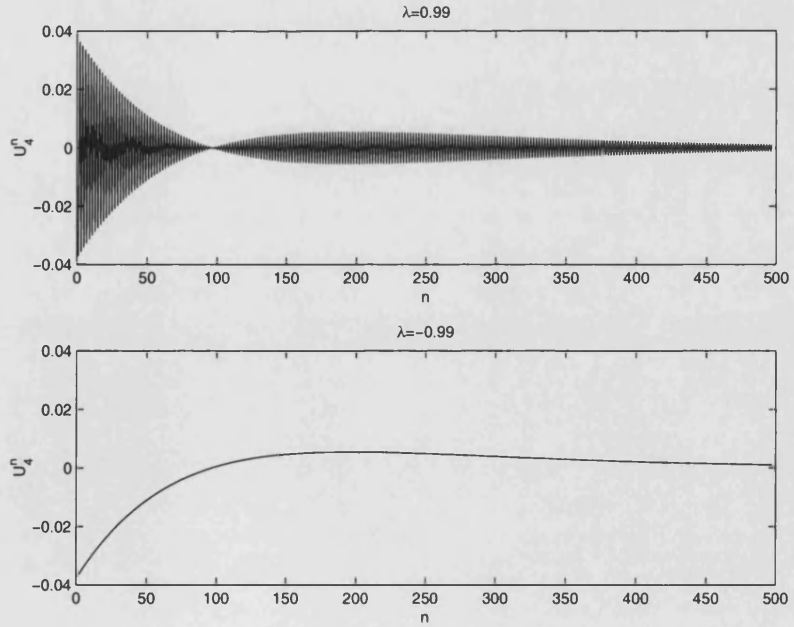

Figure B-3: U_4^n *for* $\lambda = 0.99$ *(top plot) and* $\lambda = -0.99$ *(bottom plot).*

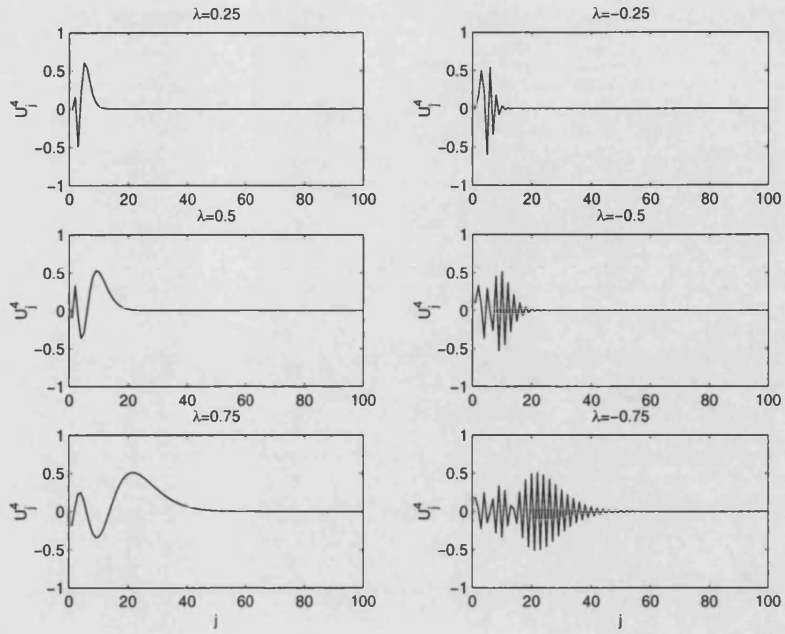

Figure B-4: U_j^4 *for various values of* λ. *In the left three plots* $\lambda > 0$ *and in the right three plots* $\lambda < 0$.

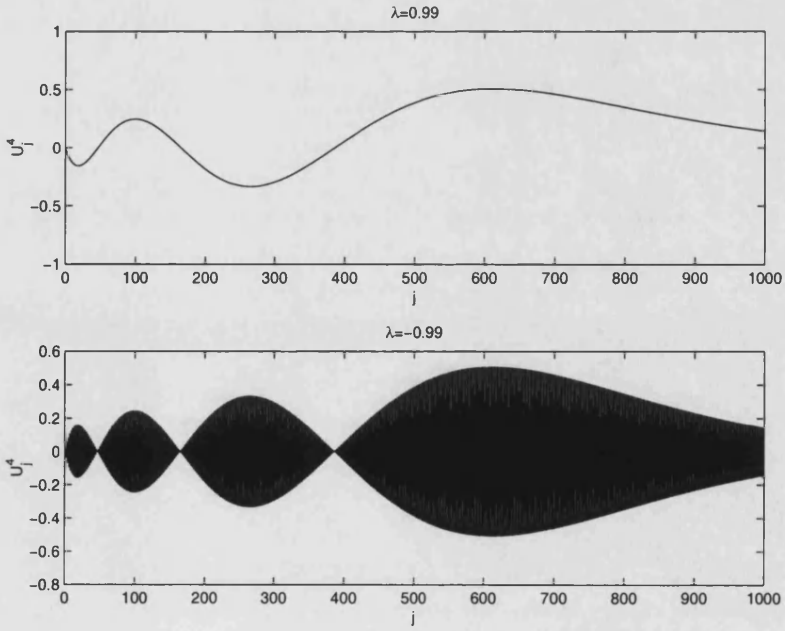

Figure B-5: U_j^4 *for* $\lambda = 0.99$ *(top plot) and* $\lambda = -0.99$ *(bottom plot).*

Definition 1. *Stability is concerned with the difference between two solutions of a numerical approximation. Since the finite difference equations being considered are linear and the difference between any two solutions has homogeneous boundary data,* ***stability****, (denoted from now on as L-R stability, see (Morton & Mayers 1994, page 157) for more details) is defined to be equivalent to establishing the following:*

$$\|\mathbf{U}^n\| \leq K\|\mathbf{U}^0\|, \tag{B.14}$$

where $\mathbf{U}^n = \{U_1^n, U_2^n, \cdots\}$ *and (B.14) holds for some norm* $\|\cdot\|$ *and some constant* K *(Morton & Mayers 1994, page 140). In this case Fourier analysis has been applied to the difference scheme and so the* ***von Neumann condition*** *is needed. It says that a necessary condition for stability is that the amplification factor* ν *satisfies* $|\nu| \leq 1 + O(\triangle t)$*. Note that a rigorous Fourier analysis is not straightforward because there is neither the whole real line as the domain nor periodic boundary conditions.*

Theorem 5. *The scheme given by (B.12) and (B.13) is L-R stable provided* $\triangle t$ *and* $\triangle x$ *satisfy the following relation:*

$$p \leq \frac{1 + \frac{1}{2}(\lambda' + \mu')}{1 + \frac{1}{2}\mu'}. \tag{B.15}$$

In order to prove (B.15), we need a general result which gives the criteria for the roots of an arbitrary quadratic, with complex coefficients, to lie inside the closed unit disc (Morton & Mayers 1994, page 147).

Lemma 11. *The roots of the polynomial* $\alpha x^2 + 2\beta x + \gamma = 0$ *with complex coefficients* α*,* β *and* γ *satisfy the condition* $|x| \leq 1$ *if and only if*

$$\textit{either} \quad |\gamma| < |\alpha| \quad \textit{and} \quad 2|\bar{\alpha}\beta - \bar{\beta}\gamma| \leq |\alpha|^2 - |\gamma|^2 \tag{B.16}$$

$$\textit{or} \quad |\gamma| = |\alpha|, \ \bar{\alpha}\beta = \bar{\beta}\gamma \quad \textit{and} \quad |\beta| \leq |\alpha|. \tag{B.17}$$

(The proof is given as an Exercise in (Morton & Mayers 1994, page 160)).

Proof. (of Theorem 5)
Suppose we substitute the Fourier modes

$$A_j^n = \hat{A}\nu^n e^{ik(j\triangle x)}, \qquad B_j^n = \hat{B}\nu^n e^{ik(j\triangle x)}, \tag{B.18}$$

into the difference equations (B.12) and (B.13). This leads to

$$\begin{aligned} (\nu - 1)(\hat{A} + \hat{B}) + 2ps(s + ic)\hat{A} &= 0 \\ (\nu - 1)\hat{B} &= \lambda'\nu\hat{A} - \mu'\nu\hat{B}, \end{aligned}$$

where $s = \sin\left(\frac{1}{2}k\triangle x\right)$, $c = \cos\left(\frac{1}{2}k\triangle x\right)$ and $\nu \equiv \nu(k)$ is the amplification factor. Suppose we set to zero the determinant of this pair of equations; we then obtain a quadratic for ν, with complex coefficients

$$\left(1+\lambda'+\mu'\right)\nu^2 - 2\left[1+\tfrac{1}{2}\lambda'+\tfrac{1}{2}\mu' - ps(1+\mu')(s+ic)\right]\nu + \left[1-2ps(s+ic)\right] = 0. \tag{B.19}$$

Lemma 11 can now be applied. So

$$|\alpha|^2 - |\gamma|^2 = 2(\lambda'+\mu') + (\lambda'+\mu')^2 - 4ps^2(p-1). \tag{B.20}$$

If $p \leq 1$ then $|\alpha|^2 - |\gamma|^2$ is obviously non-negative. If $p > 1$ then

$$|\alpha|^2 - |\gamma|^2 \geq 0 \quad \forall s^2 \quad \Longleftrightarrow \quad p(p-1) \leq \frac{1}{2}(\lambda'+\mu') + \frac{1}{4}(\lambda'+\mu')^2.$$

Substituting $p = 1+\phi$ into the above expression gives

$$\phi(1+\phi) \leq \tfrac{1}{2}(\lambda'+\mu')\left(1+\tfrac{1}{2}(\lambda'+\mu')\right),$$

and so $\phi \leq \frac{1}{2}(\lambda'+\mu')$, or

$$p \leq 1 + \tfrac{1}{2}(\lambda'+\mu'). \tag{B.21}$$

Hence the first condition in (B.16) holds provided (B.21) is true. Also

$$\begin{aligned} 2(\bar{\alpha}\beta - \bar{\beta}\gamma) &= -(\lambda'+\mu')(2+\lambda'+\mu') + 2ps^2\left[\mu'(\lambda'+\mu') - 2 + 2p(1+\mu')\right] \\ &\quad +2ipsc\mu'(2+\lambda'+\mu'). \end{aligned}$$

This can be rewritten as

$$\begin{aligned} 2(\bar{\alpha}\beta - \bar{\beta}\gamma) &= (\lambda'+\mu'+2p)\left[2p(1+\mu') - 2 - (\lambda'+\mu')\right] \\ &\quad -2pc^2\left[\mu'(\lambda'+\mu') - 2 + 2p(1+\mu')\right] + 2ipsc\mu'(2+\lambda'+\mu'). \end{aligned}$$

Finally, rearranging $|\alpha|^2 - |\gamma|^2$ gives

$$|\alpha|^2 - |\gamma|^2 = [\lambda'+\mu'-2(p-1)](\lambda'+\mu'+2p) + 4p(p-1)c^2.$$

The second condition in the statement (B.16) is equivalent to

$$\left(2|\bar{\alpha}\beta - \bar{\beta}\gamma|\right)^2 \leq \left(|\alpha|^2 - |\gamma|^2\right)^2, \tag{B.22}$$

which, in this case, becomes

$$\left\{[2p(1+\mu')-2-(\lambda'+\mu')]-\tfrac{2pc^2[\mu'(\lambda'+\mu')-2+2p(1+\mu')]}{\lambda'+\mu'+2p}\right\}^2$$
$$+\left\{\tfrac{2psc\mu'(2+\lambda'+\mu')}{\lambda'+\mu'+2p}\right\}^2 \leq \left\{[\lambda'+\mu'-2(p-1)]+\tfrac{4p(p-1)c^2}{\lambda'+\mu'+2p}\right\}^2.$$

After some manipulation this reduces to

$$\{(\mu'-\phi)(\lambda'+\mu'+2p)-c^2[\mu'(\lambda'+\mu'+2p)+2(p-1)]\}^2+4p^2s^2c^2\mu'^2(1+\phi)^2$$
$$-\{\phi(\lambda'+\mu'+2p)+2c^2(p-1)\}^2 \leq 0. \tag{B.23}$$

Let us consider the left hand side of this inequality. At $c^2=1$ the expression is zero and so the inequality (B.23) holds; at $c^2=0$ it holds provided $\phi \geq \frac{1}{2}\mu'$. Also, the derivative of this expression is positive at $c^2=1$. Hence, since $c^2 \in [0,1]$, the quadratic is always non-positive in this interval provided $\phi \geq \frac{1}{2}\mu'$, i.e.

$$p \leq \frac{1+\frac{1}{2}(\lambda'+\mu')}{1+\frac{1}{2}\mu'}.$$

This is a necessary condition and thus the proof is complete. □

Remarks

1. Suppose $\Delta t,\ \Delta x \to 0$ such that $\frac{\Delta t}{\Delta x}$ equals an arbitrary constant. Then (B.15) reduces to $p \leq 1$.

2. If Δx and Δt are fixed and λ and μ become large then the right hand side of (B.15) becomes
$$\frac{1+\frac{1}{2}(\lambda'+\mu')}{1+\frac{1}{2}\mu'} \approx \frac{\lambda+\mu}{\mu}, \tag{B.24}$$
and so the stability condition becomes $p \leq (\lambda+\mu)/\mu$. We can rewrite this as $V'\Delta t/\Delta x \leq 1$, which is precisely the condition required for the equilibrium model $a_t + V'a_x = 0$. This is to be expected since the equilibrium model arises from assuming λ and μ are large, which has been done above.

3. We will not analyse the ETIR method in any more detail since we do not want to have a stability restriction on Δt. The parameters λ and μ would have to be very large for the right hand side of (B.15) to be close to $(\lambda+\mu)/\mu$ (unless Δt is also large which will affect the accuracy of the solution). Hence we do not try to find a sufficient condition for stability.

B.3 The box-trap scheme

Theorem 6. *The box and trapezoidal scheme applied to the Linear Model written in the form (B.10) and (B.11) is given by*

$$\begin{aligned}(1+p)A_{j+1}^{n+1}+(1-p)A_j^{n+1}-(1-p)A_{j+1}^n-(1+p)A_j^n \\ +\left(B_{j+1}^{n+1}-B_{j+1}^n\right)+\left(B_j^{n+1}-B_j^n\right)=0 \qquad \text{(B.25)}\end{aligned}$$

$$B_{j+1}^{n+1}=\frac{\frac{1}{2}\lambda'}{1+\frac{1}{2}\mu'}\left(A_{j+1}^{n+1}+A_{j+1}^n\right)+\left(\frac{1-\frac{1}{2}\mu'}{1+\frac{1}{2}\mu'}\right)B_{j+1}^n. \qquad \text{(B.26)}$$

This is L-R stable (see Definition 1) for all Δx and Δt.

Proof. Suppose the modes in (B.18) are substituted into the difference equations (B.25) and (B.26). These can be divided by $\nu^n e^{ik(j\Delta x)}$ to obtain the following:

$$(\nu-1)\left(e^{ik\Delta x}+1\right)\left(\hat{A}+\hat{B}\right)+p(\nu+1)\left(e^{ik\Delta x}-1\right)\hat{A}=0, \qquad \text{(B.27)}$$

$$(\nu-1)\hat{B}=\tfrac{1}{2}(\nu+1)\left[\lambda'\hat{A}-\mu'\hat{B}\right]. \qquad \text{(B.28)}$$

We can equate to zero the determinant of the system (B.27) and (B.28) to obtain the following quadratic for ν:

$$\begin{aligned}\left[c\left(1+\tfrac{1}{2}\lambda'+\tfrac{1}{2}\mu'\right)+pis\left(1+\tfrac{1}{2}\mu'\right)\right]\nu^2+2\left(\tfrac{1}{2}\mu' pis-c\right)\nu \\ +\left[c\left(1-\tfrac{1}{2}\lambda'-\tfrac{1}{2}\mu'\right)-pis\left(1-\tfrac{1}{2}\mu'\right)\right]=0, \qquad \text{(B.29)}\end{aligned}$$

where we have used the identity

$$\frac{e^{ik\Delta x}-1}{e^{ik\Delta x}+1}=\frac{is}{c},$$

(with $s=\sin(\frac{1}{2}k\Delta x)$ and $c=\cos(\frac{1}{2}k\Delta x)$). Lemma 11 can now be applied to (B.29): if the statement (B.16) holds then the scheme is L-R stable for all Δt and Δx. Now

$$|\alpha|^2-|\gamma|^2 \;=\; 2\left(\lambda'+\mu'\right)c^2+2\mu' p^2 s^2>0,$$

and so the first condition (B.16) holds. Also

$$\begin{aligned}2|\bar{\alpha}\beta-\bar{\beta}\gamma| &= 2\left|\bar{\alpha}\left(-c+\tfrac{1}{2}p\mu' is\right)-\left(-c-\tfrac{1}{2}p\mu' is\right)\gamma\right| \\ &= 2\left|-c^2(\lambda'+\mu')+p^2\mu' s^2+2\mu' pics\right|.\end{aligned}$$

The second condition of statement (B.16) is equivalent to (B.22) which lead to

$$\left(|\alpha|^2-|\gamma|^2\right)^2-\left(2|\bar{\alpha}\beta-\bar{\beta}\gamma|\right)^2 \;=\; 16\mu'\lambda' c^2 s^2\geq 0,$$

since λ', $\mu' > 0$. Hence the box-trap scheme is L-R stable for all values of Δx and Δt and this completes the proof. □

B.4 The weighted box-trap scheme

Theorem 7. *The weighted box and trapezoidal scheme applied to the Linear Model written in the form (B.10) and (B.11) is given by*

$$(1+2\theta p)A_{j+1}^{n+1} + (1-2\theta p)A_j^{n+1} - [1-2(1-\theta)p]A_{j+1}^n - [1+2(1-\theta)p]A_j^n + (B_{j+1}^{n+1} - B_{j+1}^n) + (B_j^{n+1} - B_j^n) = 0 \tag{B.30}$$

$$B_{j+1}^{n+1} = \frac{\frac{1}{2}\lambda'}{1+\frac{1}{2}\mu'}(A_{j+1}^{n+1} + A_{j+1}^n) + \left(\frac{1-\frac{1}{2}\mu'}{1+\frac{1}{2}\mu'}\right) B_{j+1}^n. \tag{B.31}$$

This is L-R stable (see Definition 1) for all Δx and Δt provided $\theta \geq 1/2$.

Proof. Following the procedure of the proof of Theorem 2 we have a quadratic to solve for ν which is given by

$$\left[c\left(1+\tfrac{1}{2}\lambda'+\tfrac{1}{2}\mu'\right) + \theta pis\,(2+\mu')\right]\nu^2 + 2\left[\left(\tfrac{1}{2}\mu' - 2\theta + 1\right)pis - c\right]\nu + \left[c\left(1-\tfrac{1}{2}\lambda'-\tfrac{1}{2}\mu'\right) - pis(1-\theta)(2-\mu')\right] = 0. \tag{B.32}$$

We can again apply Lemma 11: if the statement (B.16) holds then the weighted box-trap scheme is L-R stable for all Δt and Δx. Now

$$\begin{aligned} |\alpha|^2 - |\gamma|^2 &= 2\left(\lambda'+\mu'\right)c^2 + \left[(2+\mu')^2\theta^2 - (2-\mu')^2(1-\theta)^2\right]p^2s^2 \\ &\geq 0 \quad \forall\, s,\, c \quad \Longleftrightarrow \quad (2+\mu')\theta \geq (2-\mu')(1-\theta), \end{aligned}$$

or

$$\theta \geq \tfrac{1}{2}\left(1-\tfrac{1}{2}\mu'\right). \tag{B.33}$$

Note that as $\Delta t \to 0$ this condition reduces to $\theta \geq \frac{1}{2}$. We also need to check that (B.22) holds. Let

$$\alpha = \alpha_r + i\alpha_i, \quad \beta = \beta_r + i\beta_i, \quad \gamma = \gamma_r + i\gamma_i,$$

then

$$\bar{\alpha} = \alpha_r - i\alpha_i, \quad \bar{\beta} = \beta_r - i\beta_i,$$

and, after some manipulation, (B.22) becomes

$$\begin{aligned} &4\left[\beta_r(\alpha_r-\gamma_r) + \beta_i(\alpha_i-\gamma_i)\right]^2 + 4\left[\beta_i(\alpha_r+\gamma_r) - \beta_r(\alpha_i+\gamma_i)\right]^2 \\ &\leq \left(\alpha_r^2-\gamma_r^2\right)^2 + 2\left(\alpha_r^2-\gamma_r^2\right)\left(\alpha_i^2-\gamma_i^2\right) + \left(\alpha_i^2-\gamma_i^2\right)^2. \end{aligned}$$

This can be written as

$$4\beta_r^2(\alpha_r-\gamma_r)^2+4\beta_i^2(\alpha_r+\gamma_r)^2-16\beta_r\beta_i(\alpha_r\gamma_i+\gamma_r\alpha_i)+4\beta_i^2(\alpha_i-\gamma_i)^2+4\beta_r^2(\alpha_i+\gamma_i)^2$$
$$\leq(\alpha_r^2-\gamma_r^2)^2+2(\alpha_r^2-\gamma_r^2)(\alpha_i^2-\gamma_i^2)+(\alpha_i^2-\gamma_i^2)^2. \quad \text{(B.34)}$$

Now

$$\begin{aligned}(\alpha_i^2-\gamma_i^2)^2-4\beta_i^2(\alpha_i-\gamma_i)^2 &= (\alpha_i-\gamma_i)^2\left[(\alpha_i+\gamma_i)^2-4\beta_i^2\right]\\ &= (\alpha_i^2-\gamma_i^2)^2p^2s^2\left\{\left[2(2\theta-1)+\mu'\right]^2-4\left(1-2\theta+\tfrac{1}{2}\mu'\right)^2\right\}\\ &\geq 0,\end{aligned}$$

if

$$2(2\theta-1)+\mu'\geq 2\left(1-2\theta+\tfrac{1}{2}\mu'\right).$$

This reduces to requiring $\theta\geq\frac{1}{2}$. Also

$$\begin{aligned}(\alpha_r^2-\gamma_r^2)^2-4\beta_r^2(\alpha_r-\gamma_r)^2 &= (\alpha_r-\gamma_r)^2\left[(\alpha_r+\gamma_r)^2-4\beta_r^2\right]\\ &= (\alpha_r^2-\gamma_r^2)^2\left[4\beta_r^2-4\beta_r^2\right]\\ &= 0.\end{aligned}$$

Hence the remaining terms in (B.34) require the following condition to hold:

$$2(\alpha_r^2-\gamma_r^2)(\alpha_i^2-\gamma_i^2)-\left[4\beta_r^2(\alpha_i+\gamma_i)^2+4\beta_I^2(\alpha_r+\gamma_r)^2-16\beta_r\beta_i(\alpha_r\gamma_i+\gamma_r\alpha_i)\right]\geq 0. \quad \text{(B.35)}$$

Now

$$\alpha_r^2-\gamma_r^2=2(\lambda'+\mu')c^2,\quad \alpha_r-\gamma_r=(\lambda'+\mu')c,\quad \alpha_r+\gamma_r=2c=-2\beta_r,$$
$$\beta_r=-c,\quad \beta_i=p\left[\tfrac{1}{2}\mu'-(2\theta-1)\right]s,$$
$$\alpha_i-\gamma_i=p\left[2+(2\theta-1)\mu'\right]s,\quad \alpha_i+\gamma_i=p\left[2(2\theta-1)+\mu'\right]s,$$

and so after extensive manipulation the condition (B.35) reduces to

$$4p^2s^2c^2\left[4\mu'\lambda'+2\mu'^2(\mu'+\lambda')(2\theta-1)\right]\geq 0$$

which holds for $\theta=\frac{1}{2}$. Hence this combined with (B.33) shows that the weighted box-trap scheme is L-R stable for all Δx and Δt provided

$$\theta\geq\tfrac{1}{2}. \quad \text{(B.36)}$$

This completes the proof. □

Appendix C

The corrected weighted box-trap scheme

In this Appendix we modify the algorithm in Section 4.4 of Chapter 4 to improve the weighted box-trap scheme for the Langmuir Model when λ and μ are large. We call this the *corrected weighted box-trap scheme.* The Langmuir Model is defined in (4.131) and (4.132) and we use data given in (4.129) and (4.130). The basic discretised equations are given in (4.133) and (4.134).

C.1 Double cell analysis for existing shock

Suppose the shock occurs in the cell (t^{m-1}, t^m) at x_j and moves to the cell (t^m, t^{m+1}) at x_{j+1}, as shown in Figure 4-16. We assume that the shock crosses the line $t = t^m$ at the point $x = x_j + \delta\triangle x$. We then have five unknowns to find: A^{m+1}_{j+1}, C^{m+1}_{j+1}, A^m_{j+1}, C^m_{j+1} and γ_{j+1}. Following the same procedure as in Section 4, the weighted box-trap scheme in the bottom cell is given by (taking into account the position of the shock)

$$C^m_{j+1/2} - \tfrac{1}{2}\big(C^{m-1}_{j+1} + C^{m-1}_j\big) + p\big[\theta A^m_{j+1} + (1-\theta)A^{m-1}_{j+1}\big] - p\big[(1-\gamma_j)A^m_j + \gamma_j A^{m-1}_j\big] = 0, \tag{C.1}$$

and

$$\begin{aligned}\big(C^m_{j+1} - A^m_{j+1}\big) - \big(C^{m-1}_{j+1} - A^{m-1}_{j+1}\big) \;=\; & \tfrac{1}{2}\lambda' A^m_{j+1}\big(B - C^m_{j+1} + A^m_{j+1}\big) - \tfrac{1}{2}\mu'\big(C^m_{j+1} - A^m_{j+1}\big) \\ & + \tfrac{1}{2}\lambda' A^{m-1}_{j+1}\big(B - C^{m-1}_{j+1} + A^{m-1}_{j+1}\big) \\ & - \tfrac{1}{2}\mu'\big(C^{m-1}_{j+1} - A^{m-1}_{j+1}\big).\end{aligned} \tag{C.2}$$

In the top cell we have

$$\begin{aligned}\tfrac{1}{2}\big(C^{m+1}_{j+1} + C^{m+1}_j\big) - C^m_{j+1/2} + C^{m-1}_j\big) + p\big[(1-\gamma_{j+1})A^{m+1}_{j+1} + \gamma_{j+1}A^m_{j+1}\big] \\ - p\big[\theta A^{m+1}_j + (1-\theta)A^m_j\big] = 0,\end{aligned} \tag{C.3}$$

and

$$\begin{aligned}(C_{j+1}^{m+1}-A_{j+1}^{m+1})-(C_{j+1}^{m}-A_{j+1}^{m}) &= \tfrac{1}{2}\lambda' A_{j+1}^{m+1}(B-C_{j+1}^{m+1}+A_{j+1}^{m+1})-\tfrac{1}{2}\mu'(C_{j+1}^{m+1}-A_{j+1}^{m+1})\\ &\quad+\tfrac{1}{2}\lambda' A_{j+1}^{m}(B-C_{j+1}^{m}+A_{j+1}^{m})-\tfrac{1}{2}\mu'(C_{j+1}^{m}-A_{j+1}^{m}).\end{aligned} \tag{C.4}$$

The quantity $C_{j+1/2}^{m}$ is specified as

$$C_{j+1/2}^{m}=\tfrac{1}{2}\delta\left[\delta C_{j+1}^{m+1}+(2-\delta)C_{j}^{m}\right]+\tfrac{1}{2}(1-\delta)\left[(1+\delta)C_{j+1}^{m}+(1-\delta)C_{j}^{m-1}\right], \tag{C.5}$$

where, by similarity of triangles

$$\delta=\frac{1-\gamma_j}{1-\gamma_j+\gamma_{j+1}}. \tag{C.6}$$

Finally, we must integrate the conservation law (4.131) across the shock

$$\tfrac{1}{2}(C_{j+1}^{m+1}+C_{j}^{m})-\tfrac{1}{2}(C_{j+1}^{m}+C_{j}^{m-1})-\tfrac{1}{2}p(\gamma_{j+1}+1-\gamma_j)\left[(A_{j+1}^{m+1}+A_{j}^{m})-(A_{j+1}^{m}+A_{j}^{m-1})\right]=0. \tag{C.7}$$

The equations (C.1), (C.2), (C.3), (C.4) and (C.7) form a nonlinear system which can be solved using a Newton iteration to give the five unknowns.

C.2 Single cell analysis for existing shock

If Δx is small enough, the shock may stay in the same cell (i.e. occur in cell (t^{m-1},t^{m}) at both x_j and x_{j+1}). Then the situation in the right diagram of Figure 4-16 holds. There are only three unknowns (A_{j+1}^{m}, C_{j+1}^{m} and γ_{j+1}) as A_{j+1}^{m+1} and C_{j+1}^{m+1} can be solved using the basic discretised equations (4.133) and (4.134). Hence, the equations are

$$\begin{aligned}&\tfrac{1}{2}(C_{j+1}^{m}+C_{j}^{m})-\tfrac{1}{2}(C_{j+1}^{m-1}+C_{j}^{m-1})+p\left[(1-\gamma_{j+1})A_{j+1}^{m}+\gamma_{j+1}A_{j+1}^{m-1}\right]\\ &\qquad -p\left[(1-\gamma_j)A_{j}^{m}+\gamma_j A_{j}^{m-1}\right]=0,\end{aligned} \tag{C.8}$$

$$\begin{aligned}(C_{j+1}^{m}-A_{j+1}^{m})-(C_{j+1}^{m-1}-A_{j+1}^{m-1}) &= \tfrac{1}{2}\lambda' A_{j+1}^{m}(B-C_{j+1}^{m}+A_{j+1}^{m})-\tfrac{1}{2}\mu'(C_{j+1}^{m}-A_{j+1}^{m})\\ &\quad+\tfrac{1}{2}\lambda' A_{j+1}^{m-1}(B-C_{j+1}^{m-1}+A_{j+1}^{m-1})\\ &\quad-\tfrac{1}{2}\mu'(C_{j+1}^{m-1}-A_{j+1}^{m-1}),\end{aligned} \tag{C.9}$$

$$\tfrac{1}{2}(C_{j+1}^{m}+C_{j}^{m})-\tfrac{1}{2}(C_{j+1}^{m-1}+C_{j}^{m-1})-\tfrac{1}{2}p(\gamma_{j+1}-\gamma_j)\left[(A_{j+1}^{m}+A_{j}^{m})-(A_{j+1}^{m-1}+A_{j}^{m-1})\right]=0. \tag{C.10}$$

C.3 Shift from the double cell to the single cell

Suppose the shock has moved to the next cell and consider both cells jointly. The flux along the right half of the top side and the whole of the right side is denoted by $G_D^]$,

$$G_D^] = \tfrac{1}{2}C_{j+1}^{m+1} + p(1-\gamma_{j+1})A_{j+1}^{m+1} + p\gamma_{j+1}A_{j+1}^{m} + p\big(\theta A_{j+1}^{m} + (1-\theta)A_{j+1}^{m-1}\big). \quad \text{(C.11)}$$

Now let $\gamma_{j+1} \to 0$. Then the shock is at the node $n = m$ at the new spatial level and we suppose that $A_{j+1}^{m} \to A_{j+1}^{m-}$. Then

$$G_D^] \longrightarrow \tfrac{1}{2}C_{j+1}^{m+1} + pA_{j+1}^{m+1} + p\theta A_{j+1}^{m-} + p(1-\theta)A_{j+1}^{m-1}. \quad \text{(C.12)}$$

Following the same procedure for $G_S^]$, which denotes the flux along the right half of the top side and the whole of the right side when the shock stays in the same cell, gives

$$G_S^] = \tfrac{1}{2}C_{j+1}^{m+1} + p\big(\theta A_{j+1}^{m+1} + (1-\theta)A_{j+1}^{m}\big) + p(1-\gamma_{j+1})A_{j+1}^{m} + p\gamma_{j+1}A_{j+1}^{m-1}, \quad \text{(C.13)}$$

and, as $\gamma_{j+1} \to 1$ (so the shock is at node $n = m$, with $A_{j+1}^{m} \to A_{j+1}^{m+}$), this becomes

$$G_S^] \longrightarrow \tfrac{1}{2}C_{j+1}^{m+1} + p\theta A_{j+1}^{m+1} + p(1-\theta)A_{j+1}^{m+} + pA_{j+1}^{m-1}. \quad \text{(C.14)}$$

In the limit these are equal if

$$(1-\theta)A_{j+1}^{m+1} - \theta A_{j+1}^{m-1} = (1-\theta)A_{j+1}^{m+} - \theta A_{j+1}^{m-}. \quad \text{(C.15)}$$

This is called the *shock jump* and is needed if, in the iteration based on the double cell, we find that $\gamma_{j+1} < 0$, and we have to apply the iteration based on the single cell. We then solve (C.8), (C.9) and (C.10) and use the values previously calculated in the double cell case as starting guesses for the Newton iteration. So we take $\theta A_{j+1}^{m} + (1-\theta)A_{j+1}^{m+} - \theta A_{j+1}^{m-}$, C_{j+1}^{m} and $1 + \gamma_{j+1}$ as the starting guesses for θA_{j+1}^{m}, C_{j+1}^{m} and γ_{j+1} respectively. In practice, the starting guess for A_{j+1}^{m} is found using (C.15) and so reduces to $A_{j+1}^{m} + \frac{1-\theta}{\theta}A_{j+1}^{m+1} - A_{j+1}^{m-1}$.

C.4 Description of the overall algorithm

In discrete form the data on the boundary data (4.130) is simply

$$A_0^n = \begin{cases} a_l, & t^n < \tau \\ a_r, & t^n > \tau, \end{cases} \quad \text{(C.16)}$$

for $n = 0, \ldots, N$ and $A_j^0 = a_l$ for $j = 0, \ldots, J$ (then define C_0^n and C_j^0). Firstly, the index $n = m - 1$ is found, which is the location of the shock at level $j = 0$. Then set

$$\gamma_0 = \frac{\tau - t^{m-1}}{\Delta t}. \tag{C.17}$$

Assume that τ is not a nodal value and so $\gamma_0 \in (0, 1)$. For each level j the values A_{j+1}^{n+1} and C_{j+1}^{n+1} are found by solving (4.133) and (4.134) for $n = 0, \ldots, m - 2$.

We are now at the point where we must modify the box scheme to allow for the shock. Assume the cell goes into the next cell at the new spatial level. Then solve the five nonlinear equations (C.1), (C.2), (C.3), (C.4) and (C.7) using Newton's method. From the theory of shock waves in nonlinear conservation laws the shock speed is $[Va]/[c]$ and so we approximate this at level j by $V(A_j^{m-1} - A_j^m)/(C_j^{m-1} - C_j^m)$. This suggests that a reasonable starting guess for γ_{j+1} is $-(1 - \gamma_j) + (C_j^{m-1} - C_j^m)/p(A_j^{m-1} - A_j^m)$, where p is the CFL number. Starting guesses for A_{j+1}^{m+1}, C_{j+1}^{m+1}, A_{j+1}^m and C_{j+1}^m are given by A_j^m, C_j^m, A_j^{m-1} and C_j^{m-1} respectively. Hence the algorithm is as follows:

1. solve (4.133) and (4.134) for $n = 0, \ldots, m - 2$.

2. iterate to find A_{j+1}^{m+1}, C_{j+1}^{m+1}, A_{j+1}^m, C_{j+1}^m and γ_{j+1} by solving (C.1), (C.2), (C.3), (C.4) and (C.7) with starting guesses A_j^m, C_j^m, A_j^{m-1}, C_j^{m-1} and $-(1 - \gamma_j) + (C_j^{m-1} - C_j^m)/p(A_j^{m-1} - A_j^m)$.

3. if $\gamma_{j+1} > 0$ find A_{j+1}^{n+1} and C_{j+1}^{n+1} for $n = m + 1, \ldots, N - 1$ using (4.133) and (4.134). Set $m = m + 1$ (since the shock is now in the next cell) and move to the next spatial level (i.e. go to step 1.). Otherwise, change A_{j+1}^m by the shock jump defined in (C.15), i.e. set

$$\bar{A}_{j+1}^m = A_{j+1}^m + \frac{1 - \theta}{\theta} A_{j+1}^{m+1} - A_{j+1}^{m-1}, \tag{C.18}$$

   and, since $\gamma_{j+1} < 0$, also set $\bar{\gamma}_{j+1} = 1 + \gamma_{j+1}$. The values A_{j+1}^{m+1}, A_{j+1}^{m-1} and γ_{j+1} are those calculated in step 1. Finally, the starting guess for C_{j+1}^m is simply $\bar{C}_{j+1}^m = C_{j+1}^m$ which was also calculated in step 1. Go to step 4.

4. solve (C.8), (C.9) and (C.10) using Newton's method to re-calculate A_{j+1}^m, C_{j+1}^m and γ_{j+1}. Take $\bar{A}_{j+1}^m$, $\bar{C}_{j+1}^m$ and $\bar{\gamma}_{j+1}$ as starting guesses. Finally, use (4.133) and (4.134) to re-calculate A_{j+1}^{m+1} and C_{j+1}^{m+1} and then to find A_{j+1}^{n+1} and C_{j+1}^{n+1} for $n = m + 1, \ldots, N - 1$. Move to the next time level (i.e. go to step 1.).

We will call this the **corrected weighted box-trap scheme**. Note that (4.133) and (4.134) leads to a quadratic to solve for A_{j+1}^{n+1}; we take the positive root.

Bibliography

Abbott, M. B. (1979), *Computational Hydraulics: Elements of the Theory of Free Surface Flows*, Pitman Publishing Limited.

Ablowitz, M. J. & Fokas, A. S. (1997), *Complex Variables: Introduction and Applications*, Cambridge University Press.

Abramowitz, M. & Stegun, I. (1965), *Handbook of Mathematical Functions*, Dover Publications, New York.

Barry, D. A., Bajracharya, K. & Miller, C. T. (1996), 'Alternative split-operator approach for solving chemical reaction/groundwater transport models', *Advances in Water Resources* **19**(5), 261–275.

Barry, D. A., Miller, C. T., Culligan, P. J. & Bajracharya, K. (1997), 'Analysis of split operator methods for nonlinear and multispecies groundwater chemical transport models', *Mathematics and Computers in Simulation* **43**, 331–341.

Bereux, F. & Sainsaulieu, L. (1997), 'A Roe-type Riemann solver for hyperbolic systems with relaxation based on time-dependent wave decomposition', *Numerische Mathematik* **77**, 143–185.

Botchorishvili, R., Perthame, B. & Vasseur, A. (2003), 'Equilibrium Schemes for Scalar Conservation Laws with Stiff Sources', *Mathematics of Computation* **72**(241), 131–157.

Budd, C. J., Carey, C. M. M., Graham, I. G. & Spence, A. (1997), Coupling transport and chemistry in groundwater flow: A study of asympotic and numerical solution methods for a model problem. Internal report.

Caflisch, R. E., Jin, S. & Russo, G. (1997), 'Uniformly Accurate Schemes for Hyperolic Systems with Relaxation', *SIAM Journal on Numerical Analysis* **34**(1), 246–281.

Cecchi, M. M., Redivo-Zaglia, M. & Russo, G. (1996), 'Extrapolation methods for hyperbolic systems with relaxation', *Journal of Computational and Applied Mathematics* **66**, 359–375.

Chen, G.-Q., Levermore, C. D. & Liu, T.-P. (1994), 'Hyperbolic Conservation Laws with Stiff Relaxation Terms and Entropy', *Communications on Pure and Applied Mathematics* **XLVII**, 787–830.

Colella, P., Majda, A. & Roytburd, V. (1986), 'Theoretical and numerical structure for reacting shock waves', *SIAM Journal on Scientific and Statistical Computing* **7**, 1059–1080.

Courant, J. (1934), *Differential and Integral Calculus*, Blackie and Son Ltd.

Cunge, J. A. & Holly Jr, F. M. Verwey, A. (1980), *Practical Aspects of Computational River Hydraulics*, Pitman Publishing Limited.

DuChateau, P. & Zachmann, D. (1989), *Applied Partial Differential Equations*, Harper and Row Publishers, Inc.

Forsythe, G. E. & Wasow, W. R. (1960), *Finite-Difference Methods for Partial Differential Equations*, John Wiley and Sons, Inc., New York.

Friedly, J. C. & Rubin, J. (1992), 'Solute Transport With Multiple Equilibrium-Controlled or Kinetically Controlled Chemical Reactions', *Water Resources Research* **28**(6), 1935–1953.

Garabedian, P. (1964), *Partial Differential Equations*, John Wiley and Sons, Inc.

Grindrod, P. (1991), *Patterns and Waves: The Theory and Applications of Reaction-Diffusion Equations*, Clarendon Press, Oxford.

Guenther, R. B. & Lee, J. W. (1988), *Partial Differential Equations of Mathematical Physics and Integral Equations*, Prentice Hall, New Jersey.

Herzer, J. & Kinzelback, W. (1989), 'Coupling of Transport an Chemical Processes in Numerical Transport Models', *Geoderma* **44**, 115–127.

Holmes, M. H. (1995), *Introduction to Perturbation Methods*, Springer-Verlag, Inc., New York.

Jeffreys, H. & Jeffreys, B. (1972), *Methods of Mathematical Physics*, Cambridge University Press.

Jennings, A. A., Kirkner, D. J. & Theis, T. L. (1982), 'Multicomponent Equilibrium Chemistry in Groundwater Quality Models', *Water Resources Research* **18**(4), 1089–1096.

Jin, S. & Levermore, C. D. (1996), 'Numerical Schemes for Hyperbolic Conservation Laws with Stiff Relaxation Terms', *Journal of Computational Physics* **126**, 449–467.

LeVeque, R. J. (1992), *Numerical Methods for Conservation Laws*, Birkhauser Verlag.

LeVeque, R. J. & Yee, H. C. (1990), 'A Study of Numerical Methods for Hyperbolic Conservation Laws with Stiff Source Terms', *Journal of Computational Physics* **86**, 187–210.

Lighthill, M. J. & Whitham, G. B. (1955), 'On kinematic waves: I. Flood movement in long rivers; II. Theory of traffic flow on long crowded roads', *Proceedings of the Royal Society, A* **229**, 281–245.

Liu, T.-P. (1987), 'Hyperbolic Conservation Laws with Relaxation', *Communications in Mathematical Physics* **108**, 153–175.

Mackenzie, J. A. (1998), 'The efficient generation of simple two-dimensional adaptive grids', *SIAM Journal on Scientific Computing* **19**(4), 1340–1365.

McOwen, R. C. (1995), *Partial Differential Equations, Methods and Applications*, Prentice Hall, New Jersey.

Mitchell, S. L., Morton, K. W. & Spence, A. (2003*a*), A comparison of the box scheme and splitting methods for coupled transport and chemistry in groundwater flow. In preparation.

Mitchell, S. L., Morton, K. W. & Spence, A. (2003*b*), Modelling reactive transport with the box scheme. In preparation.

Molz, F. J., Widdowson, M. A. & Benefield, L. D. (1986), 'Simulation of microbial dynamics coupled to nutrient and oxygen transport in porous media', *Water Resources Research* **22**(8), 1207–1216.

Morton, K. W. (1996), *Numerical Solution of Convection-Diffusion Problems*, Chapman and Hall, London.

Morton, K. W. & Mayers, D. F. (1994), *Numerical Solution of Partial Differential Equations*, Cambridge University Press.

Papalexandris, M. V., Leonard, A. & Dimotakis, P. E. (1997), 'Unsplit Schemes for Hyperbolic Conservation Laws with Source Terms in One Space Dimension', *Journal of Computational Physics* **134**, 41–61.

Pember, R. B. (1993*a*), 'Numerical Methods for Hyperbolic Conservation Laws with Stiff Relaxation I. Spurious Solutions', *SIAM Journal on Applied Mathematics* **53**(5), 1293–1330.

Pember, R. B. (1993*b*), 'Numerical Methods for Hyperbolic Conservation Laws with Stiff Relaxation II. Higher-order Godunov methods', *SIAM Journal on Scientific Computation* **14**(4), 825–859.

Rhee, H.-K., Aris, R. & Amundson, N. R. (1986), *First-Order Partial Differential Equations, Volume I, Theory and Application of Single Equations*, Prentice Hall.

Richtymer, R. D. & Morton, K. W. (1967), *Difference Methods for Initial-Value Problems*, John Wiley and Sons, Inc., New York.

Rubin, J. (1983), 'Transport of reacting solutes in porous media: relation between mathematical nature of problem formulation and chemical nature of reactions', *Water Resources Research* **9**(5), 1231–1252.

Spiegel, M. R. (1959), *Schaum's Outline of Theory and Problems of Vector Analysis*, Schaum Publishing Co.

Spiegel, M. R. (1971), *Schaum's Outline of Theory and Problems of Calculus of Finite Differences and Difference Equations*, McGraw-Hill Book Company.

Strang, G. (1968), 'On the construction and comparison of difference schemes', *SIAM Journal on Numerical Analysis* **5**, 506–517.

Trefethen, L. N. (1982), 'Group velocity in finite difference schemes', *SIAM Review* **24**(2), 113–136.

Walter, A. L., Frind, E. O., Blowes, D. W., Ptacek, C. J. & Molson, J. W. (1994*a*), 'Modeling of multicomponent reactive transport in groundwater 1. Model development and evaluation', *Water Resources Research* **30**(11), 3117–3148.

Walter, A. L., Frind, E. O., Blowes, D. W., Ptacek, C. J. & Molson, J. W. (1994*b*), 'Modeling of multicomponent reactive transport in groundwater 2. Metal mobility in aquifers impacted by acidic mine tailings discharge', *Water Resources Research* **30**(11), 3149–3158.

Walter, W. (1998), *Ordinary Differential Equations, Graduate Texts in Mathematics*, Springer-Verlag Inc., New York.

Warming, R. F. & Hyett, B. J. (1974), 'The Modified Equation Approach to the Stability and Accuracy Analysis of Finite-Difference Methods', *Journal of Computational Physics* **14**, 159–179.

Wheeler, M. F. & Dawson, C. N. (1988), An operator-splitting method for advection-diffusion-reaction problems, *in* 'The Mathematics of Finite Elements and Applications VI', Academic Press Ltd.

Whitham, G. B. (1974), *Linear and Nonlinear Waves*, John Wiley and Sons, Inc., New York.

Yeh, G. T. & Gwo, J. P. (1990), A Lagrangian-Eulerian Approach to Modeling Multicomponent Reactive Transport, *in* 'Proceedings of VIII International Conference on Computational Methods in Water Resources', Venice, Italy, pp. 419–427.

Zysset, A., Stauffer, F. & Dracos, T. (1994), 'Modelling of chemically reactive groundwater transport', *Water Resources Research* **30**(7), 2217–2228.